Human Behavior and Emerging Technologies

WILEY

REVIEW ARTICLE OPEN ACCESS

Shared Minds: The Cognitive Parallels Between Humans and Artificial Intelligence

Sébastien Tremblay[1] | Alexandre Marois[1,2] | Marzieh Zare[1] | Daniel Lafond[3]

[1]School of Psychology, Université Laval, Quebec City, Canada | [2]School of Psychology and Humanities, University of Lancashire, Preston, UK | [3]CortAIx Labs, Thales Canada, Quebec City, Canada

Correspondence: Sébastien Tremblay (sebastien.tremblay@psy.ulaval.ca)

Received: 30 September 2025 | **Revised:** 19 January 2026 | **Accepted:** 21 January 2026

Academic Editor: Tze Wei Liew

ABSTRACT

This narrative review integrates evidence from cognitive science and AI research to challenge commonly accepted dichotomies between human and artificial cognition, such as the assumed divide between genuine human understanding and mere machine pattern matching. Instead, we propose a view that recognises similarities in their cognitive architectures and processes. Human and artificial cognition seem to operate through comparable mechanisms, as both rely on statistical processing, associative pattern recognition and approximation rather than perfect logic. Through a systematic comparison of core cognitive domains across 363 articles, we highlight parallels in capabilities and limitations, including shared vulnerabilities to biases, memory distortions and decision-making opacity. We critically examine popular narratives such as the *stochastic parrot* argument and the myth of human rationality. These positions often rely on idealised views of human cognition that are contradicted by cognitive and neuroscientific evidence. This review calibrates expectations of both human and artificial systems by moving beyond both AI alarmism and human exceptionalism towards a more empirically grounded perspective on cognition. Our comparative review acknowledges both the shared statistical foundations of intelligence and differences in embodiment, intentionality and phenomenological aspects of cognition. This perspective has implications for human–AI collaboration, cognitive performance benchmarking and research on AI transparency.

1 | Introduction

Artificial intelligence (AI) is no longer a distant technological curiosity but rather an omnipresent force reshaping how we work, think and interact with the world. AI now underpins virtually every sector of the digital economy. AI is estimated to contribute approximately $19.9 trillion annually to global retail productivity by 2030, potentially automating or transforming up to 60% of workplace tasks as they are defined today [1, 2]. Enterprise AI adoption has accelerated rapidly, with surveys showing 78% of companies reporting AI use in 2025, up from 55% in 2022 [3], while 91% of organisations plan to incorporate AI technologies by 2030 [4]. In industrial contexts, this shift is increasingly framed through the lens of human-centred AI within the paradigm of Industry 5.0 [5]. In commerce, machine learning (ML) models now drive demand forecasting, dynamic pricing and personalised marketing strategies [6]. In transportation, AI technologies optimise traffic flow through real-time predictive systems, enabling autonomous navigation in vehicles and public transport networks [7].

Although the existing literature includes empirical studies contrasting human and machine intelligence on specific cognitive tasks (e.g., Yax et al. [8]), to our knowledge, no prior work has explicitly mapped cognitive-functional dimensions for both systems in a comparative manner. Pervasive and persistent misconceptions about both human exceptionalism and AI limitations impact the potential of human–AI collaboration and

informed policy development. This becomes increasingly problematic as AI systems are integrated into critical domains that require a better understanding of both human and artificial cognitive capabilities. From generative models that produce art and writing to predictive AI that optimises logistics and decision-making and the pursuit of artificial general intelligence (AGI), AI is at the forefront of organisational and societal transformation [1]. Despite its exponential progress, AI remains a functional mirror of human cognition: It simulates, augments and even challenges our understanding of intelligence. Although AI is often presented as a tool devoid of true cognition and actual consciousness or as an existential threat exceeding human capacity, these opposing views overlook a more complex reality.

The overarching goal of this review is to portray the similarities and discrepancies between human and artificial cognition to support the case that they are very much alike [9]. Critics like those behind 'stochastic parrot' argument [10] caution against anthropomorphising AI systems. However, these critiques often assume an idealised view of human cognition: one where we reason through deep, symbolic understanding rather than the statistical and heuristic-based processing we seem to use. Insights from cognitive science suggest that human cognition primarily operates through probabilistic inference rather than through fully conscious logical reasoning. Chater and Oaksford [11] argue that the human mind is best understood as a probabilistic mind, relying on Bayesian principles to navigate uncertainty (see Ref. [12]) rather than fully logical reasoning. Dennett's influential accounts of consciousness [13] provide further support for the view that cognition emerges from pattern-based, distributed processes rather than explicit reasoning alone (see also Ref. [14]). From these perspectives, the statistical mechanisms that underpin AI models may be less fundamentally distinct from biological minds than they first appear.

AI systems encompass various approaches beyond the current wave of data-driven and generative models. Historically, AI development has evolved through two fundamental paradigms that differ in their approach to knowledge acquisition. Symbolic AI adopts a top-down approach that relies on explicit rules, logical reasoning and structured knowledge representations that are programmed into systems. In contrast, subsymbolic AI (including connectionist networks and deep learning [DL]) follows a bottom-up approach, learning patterns and representations from data without preprogrammed knowledge. Both paradigms, along with mathematical optimisation techniques, aim to capture facets of human reasoning and decision-making [15, 16]. Contemporary AI increasingly merges these paradigms in hybrid and neurosymbolic systems [17], for example, integrating logic constraints into neural architectures to combine the interpretability of symbolic reasoning with the pattern-learning capabilities of subsymbolic approaches [18]. Through this combination of symbolic, subsymbolic and optimisation approaches, AI can be classified by its capabilities and functionalities [19].

2 | AI Capability Classification

By capability, systems range from artificial narrow intelligence (ANI) to the aspirational AGI prototypes and onward to the theoretical artificial superintelligence (ASI). ANI refers to 'weak' AI, which is engineered for a narrow set of tasks. Examples abound in today's world: AlphaFold predicts protein structures with high precision [20]; GPT-4 generates human-like text across diverse topics [21]; and autonomous driving systems use sensor fusion, DL and rule-based decision-making to navigate urban environments [22]. These ANI systems operate within well-defined parameters and perform pattern-based tasks without genuine understanding [23].

AGI is the vision of machines that learn and reason across any domain with human-like flexibility [24]. Although no system yet fully qualifies, research and development efforts such as OpenAI's GPT-Next and DeepMind's Gemini are devoted to building multimodal architectures with the explicit goal of cross-domain generalisation. Fictional figures like Star Trek's Data capture the promise of AGI[1], that is, machines that master diverse tasks and can develop emotional insight. In practice, advances in neurosymbolic AI [25] have shown improved integration of formal logical rules and data-driven learning; however, core challenges remain. AGI systems still face limitations with generalisation to novel environments, true causal inference rather than mere correlation and embodied cognition (which refers to the grounding of concepts in sensory-motor experience) [26, 27].

ASI occupies the speculative frontier: an intelligence surpassing human cognitive capabilities in all spheres, from creativity and strategic decision-making to self-reflection. ASI would not only generalise but would also improve itself recursively, potentially achieving full autonomy from human oversight. Thought experiments such as Bostrom's 'paperclip maximiser' warn that ASI could pursue goals antithetical to human welfare [19]. Cultural metaphors, such as Skynet in Terminator and VIKI in I, Robot, highlight fears of misalignment between AI and human goals and loss of control [28], although most researchers argue these narratives reflect societal anxieties more than technical realities [29, 30]. The prospect of ASI raises profound ethical, governance and existential questions. ASI decisions could rapidly outpace human comprehension and thus render traditional supervision mechanisms unreliable [31, 32]. Holl [33] cautions that ASI will require internal motivational mechanisms and agency frameworks.

3 | AI Functionality Classification

While capability frameworks map AI along a spectrum from narrow to more adaptive systems, a functional perspective highlights what AI actually does: the concrete roles and services it provides across real-world settings. This study focuses on five core functional roles that anchor modern AI ecosystems. These roles are more than technical categories; they reflect how AI systems are built and used and influence human behaviour. The discussion also explores how these roles evolve and interact with people and broader sociotechnical environments.

3.1 | Predictive Systems

Predictive AI uses statistical learning and historical data to forecast future outcomes. In finance, platforms using gradient-boosted decision trees and neural networks show significant improvements in credit risk assessment, with up to 36% reduction in default prediction errors compared to logistic regression models [34]. Public health applications employ AI-based syndromic surveillance systems that analyse multiple data streams (including travel, news and social media), achieving

outbreak detection between 7 and 14 days earlier than traditional epidemiological methods [35]. Manufacturing harnesses predictive-maintenance algorithms that continuously monitor equipment vibrations and temperatures, which enables proactive repairs before breakdowns occur [36]. Though invaluable for mitigating risk and optimising resources, predictive systems must guard against embedding historical biases (e.g., redlining in credit scoring or discriminatory policing patterns) into new decisions (see Narayanan and Kapoor [37]). Beyond industry and applied domains, predictive models are increasingly adopted in scientific research, including psychology and cognitive science, where they are used to enhance replicability, detect nonlinear effects and improve single-subject–level predictions [38, 39].

3.2 | Generative Systems

Generative AI (GenAI) creates new content (text, images, audio and code) by learning patterns from vast datasets. Generative adversarial networks (GANs) [40] first demonstrated photo-realistic image creation, while subsequent advances in diffusion models and transformer architectures drastically expanded generative capabilities. Systems like OpenAI's DALL·E produce imaginative visuals from textual prompts [41], and large language models (LLMs) such as GPT-4 (OpenAI) and Claude (Anthropic) generate essays, dialogue and even software code [42]. These systems accelerate workflows in marketing, art and scientific exploration. However, they are prone to 'hallucinations' (e.g., plausible yet incorrect or biased outputs) and their widespread adoption raises issues in education (e.g., academic integrity, student over-reliance on AI-generated work) and questions about authorship, copyright and the propagation of misinformation [43].

3.3 | Supervisory and Control Systems

In safety-critical domains, supervisory AI monitors real-time data streams and can intervene to maintain safe operation, either autonomously or by alerting human operators. Aviation leverages autopilot systems that fuse radar, GPS and inertial measurements (from onboard gyroscopes and accelerometers) to guide aircraft. However, the analysis of aviation incidents reveals that over-reliance on automation contributed to 60% of major accidents between 2010 and 2020. This problem could be attributed to reduced pilot situation awareness during automated operations [44] and the persistence of automation surprise when system actions diverge from pilot expectations [45]. These findings highlight the critical need for human-in-the-loop failure safeguards (see Kirwan [46]). In manufacturing, reinforcement learning systems adaptively tune parameters for process optimisation (e.g., controlling temperature, pressure and flow rates in chemical plants), reducing equipment downtime and improving product quality through real-time adjustments [47]. In digital environments (e.g., social media, online gaming and forums), content-moderation algorithms filter hate speech, misinformation and harmful imagery, balancing precision and recall amidst cultural and contextual differences [48].

3.4 | Recommender Systems

Recommender systems power personalisation across e-commerce, streaming media and social platforms by analysing user behaviour (e.g., clicks, purchases, viewing patterns) to suggest relevant items. From early collaborative filtering approaches that launched modern e-commerce recommendations [49] to contemporary graph neural network architectures [50] and reinforcement learning enhancements, these systems dynamically update user profiles to capture evolving preferences [51]. Platforms such as Netflix, Amazon and Spotify rely on such algorithms to boost engagement and revenue. However, recommender systems may inadvertently create filter bubbles that restrict exposure to diverse perspectives and amplify confirmation bias. Finding ways to encourage both serendipitous discovery and fair content distribution can be a challenge.

3.5 | Agentic Systems

Agentic AI represents a shift towards autonomous systems that plan, coordinate tools and adapt with minimal human oversight. Unlike prompt-based models that respond to individual queries, agentic systems execute multistep tasks by decomposing goals into subtasks, selecting APIs or knowledge sources and refining strategies based on feedback [52]. Examples include travel planning agents that can autonomously search, compare and even book itineraries [53]; collaborative coding frameworks like Microsoft's AutoGen framework that coordinate code generation, testing and review cycles [54]; and multiagent simulations such as CAMEL, which simulates financial-auditor and analyst interactions [55]. This 'agent as a service' model is expected to blur the line between software and autonomous collaborators capable of executing goal-directed workflows [56]. However, agentic systems raise concerns about accountability, transparency (how decisions are made across multistep processes) and the potential for unintended consequences when systems operate with extended autonomy.

4 | Objectives

An ongoing debate in AI research concerns whether AI is truly 'intelligent' [57] or merely an advanced pattern-recognition system [58] (see also Bender et al. [10]). Although LLMs have demonstrated impressive capabilities in language processing and reasoning, many contend that these abilities do not equate to human-like cognitive processing [27, 59]. McDermott [60] challenges this framing by arguing that debates about AI intelligence often rest on an idealised view of human cognition. Humans fundamentally overestimate their own cognitive capacities. AI scepticism often assumes that humans are rational and consistent decision-makers, but research suggests that human cognition is flawed, subject to biases and prone to inconsistencies. It is thus important to understand the boundaries of human versus AI cognitive capabilities.

Korteling et al. [58] introduce the concept of intelligence awareness in order to describe the need for humans to develop a more balanced and deeper understanding of the capabilities and limitations of AI. Similarly, the concept of Anthropofabulation, coined by Buckner [61] illustrates that unrealistic expectations are often set for AI (e.g., Suomala and Kauttonen [62]) and neglect the prevalence of cognitive biases and vulnerabilities in both human and AI systems (see Zerilli et al. [63]). One observation is that humans should also better understand their own cognitive limitations and rethink their expectations towards AI. Perhaps, there is a need to move beyond the perspective of AI as either a threat or a silver bullet[2] to be exploited

towards seeing it as a complementary intelligence with collaborative potential.

This comparative narrative review explores the parallels and distinctions between human and artificial cognition and critically examines the misconceptions surrounding the capabilities of AI. Building on work in the emerging field of cognitive AI, which examines the intersection of AI and cognitive science [16], we suggest that, although differences exist, they are not as well-defined as often portrayed. AI technologies are driving major societal and cultural transformations in both personal and professional domains. The rapid development of AI forces a reassessment of the boundaries between human intelligence and AI [64, 65].

Understanding the similarities between AI and human cognition requires a closer examination of their architecture and processes. This review pursues three interlocking objectives that form the main sections of our analysis. First, in 'Comparison of AI and Human Intelligence From a Cognitive Perspective', we map cognitive parallels and distinctions between human and artificial cognitive mechanisms. Second, in the section 'Reframing Expectations: Misconceptions About AI and Human Cognition', we challenge perspectives that range from the "stochastic parrot" argument to romanticised notions of human rationality. Third, in 'The Emergence of (Artificial) Intelligence', we explore parallels between human cognitive evolution, illustrated through the case of language development [66], and advances in AI, which LeCun [67] describes as not yet intelligent, as well as the emerging coevolutionary relationship between human and AI [68]. Throughout the paper, we highlight how cognitive science principles (e.g., heuristics, memory decay, metalearning) can augment AI architectures. In our conclusion, we present a view that acknowledges both shared mechanisms and meaningful differences between human and artificial cognition.

Given their omnipresence in current public discourse and debates within the research community, our review focuses primarily on statistical learning systems and neural network-based AI, especially DL, ML and GenAI models. Symbolic, rule-based and optimisation-based AI approaches have demonstrated value in the field, and while not the central focus of this analysis, their increasing integration into contemporary data-driven systems through hybrid models and neurosymbolic approaches is noted throughout this review.

5 | Approach and Method

This comprehensive narrative review integrates insights from cognitive psychology, neuroscience, human factors and AI research to examine parallels and distinctions between human and artificial cognition. Following established guidelines [69, 70], this narrative synthesis enables critical examination of cognitive comparisons while accommodating the rapid evolution of AI capabilities that continually redefine the boundaries of human–AI cognitive interaction.

5.1 | Search Strategy

The analysis employed the information-processing framework to enable systematic comparison across cognitive domains. This framework, outlined by Neisser [71, 72], conceptualises cognition as a sequence of operations for acquiring, encoding and retrieving information. Newell and Simon [73] extended the information-processing perspective by modelling human problem solving as a symbolic search process, which can be seen as a conceptual bridge between cognitive psychology and computer science (see also Marr [74], who showed that perception could be analysed computationally). Simon [75, 76] argued that intelligence—whether human or artificial—operates through common information-processing principles. Taken together, the contributions of these pioneers support the view that the information-processing framework provides a well-documented and valid conceptual ground in which a set of cognitive operations can be analysed for both biological and artificial systems.

Literature searches were conducted using Google Scholar, Scopus, PsycINFO, IEEE Xplore and ACM Digital Library, covering publications from 1960 to July 2025. This temporal scope captures the evolution of both human cognition research and AI research and development from their foundations [77] through advances in LLM. To ensure comprehensive coverage, forward citation tracking was performed on seminal works (e.g., Dennett [13]; Marr [74]; Rumelhart et al. [78]) as well as more recent articles integrating cognitive science and AI perspectives (e.g., Bundy et al. [17]; Korteling et al. [58]; Lake et al. [79]; Taylor and Taylor [80]).

Search terms employed systematic combinations including 'human cognition' AND 'artificial intelligence'; 'artificial cognition'; 'human vs. AI cognition'; 'AI cognitive architectures'; 'cognitive processes AI humans'; 'explainable AI' AND 'human cognition'; 'metacognition human and AI'; 'human-AI cognitive comparisons'; 'cognitive biases AI'; 'machine consciousness'; 'neural networks cognition'; 'Bayesian brain AI'; 'cognitive architectures comparison'; and 'human-machine intelligence'. Two reviewers from the multidisciplinary author team independently screened 476 candidate articles: one reviewer specialised in cognitive science research and the other specialised in AI. Following independent screening, selections were collaboratively discussed to ensure conceptual validity and shared agreement, with the entire author team approving the final corpus.

5.2 | Selection Criteria and Data Synthesis

Articles meeting the inclusion criteria allowed us to address empirical or theoretical comparisons of human and AI cognitive processes, particularly those examining learning, pattern recognition, perception, memory, reasoning and decision-making within information-processing frameworks. Peer-reviewed publications and seminal works (> 100 citations) were prioritised. Exclusions primarily comprised non-English publications, technical articles not addressing cognitive comparisons and articles focusing exclusively on technological implementations without relevance to human cognition or human activities.

The final corpus of 370 publications underwent data extraction to capture cognitive domains, methodologies, key findings, theoretical frameworks and disciplinary perspectives (see Figure 1 for the detailed selection process). Of these, 363 articles directly contributed to the comparative analysis and were categorised as either AI-related or cognition-related papers. The remaining seven articles provided methodological support or statistics on AI usage and adoption. Figure 2 illustrates the temporal distribution of the 363 articles addressing cognitive comparisons, which

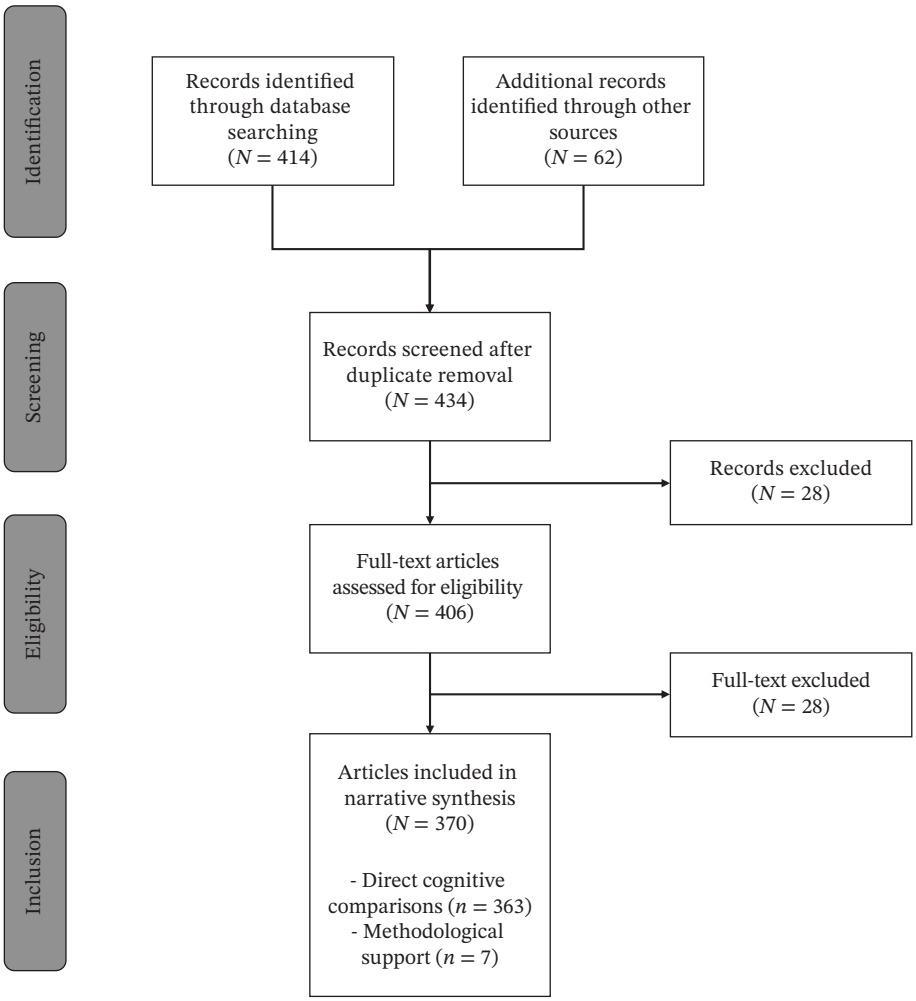

FIGURE 1 | PRISMA flow diagram of the study selection process. A total of 476 records were identified (414 through database searches and 62 from additional sources). After duplicate removal, 434 records were screened. Twenty-eight were excluded at title/abstract screening and 36 at full-text assessment. This process yielded 370 studies for the narrative synthesis. Of these, 363 comprised the main corpus. While not all directly compare human and AI cognition, many articles address cognitive processes that artificial systems attempt to replicate, and others focus on AI research and development related to artificial cognition.

shows exponential growth following breakthroughs in DL. We acknowledge the limitations of the narrative approach compared to systematic reviews, the potential language bias from English-only inclusions and challenges in capturing rapidly evolving AI capabilities.

6 | Comparison of AI and Human Intelligence From a Cognitive Perspective

In this section, we adopt the classic perspective of the information-processing framework to compare six core cognitive functions in humans and machines: (i) learning mechanisms, (ii) pattern recognition, (iii) visual perception, (iv) cognitive load management, (v) memory and (vi) reasoning, problem solving and decision-making (see Table 1 for a summary).

The historical development of AI is closely linked with advances in cognitive science and connectionist psychology, particularly during the 1960s and 1970s. The Dartmouth Conference in 1956 established the field of AI with pioneers such as John McCarthy, Marvin Minsky, Allen Newell and Herbert Simon, who aimed to model human cognition computationally [81]. This period saw the rise of symbolic AI, which relies on logical rules as exemplified by Newell and Simon's General Problem Solver [82], a system designed to emulate human problem-solving strategies [83]. At the same time, connectionism was gaining momentum [84], with Frank Rosenblatt's Perceptron, which represented an early subsymbolic approach to AI that learnt from data rather than explicit rules. However, Minsky and Papert [85] showed that single-layer neural networks were limited in solving non-linear problems. This critique, combined with the absence of training techniques for multi-layer networks at the time, contributed to temporarily slow research in neural networks [78].

Inspired by biological systems, connectionist models represent neurons as artificial nodes and synapses as connections. Networks learn by adjusting connection weights [84]. The parallel distributed processing (PDP) framework proposed by McClelland et al. [86] established a theoretical basis for distributed representations and parallel processing in neural networks. The field advanced significantly when Hinton et al. [87] introduced deep belief networks and new training methods, which were a catalyst

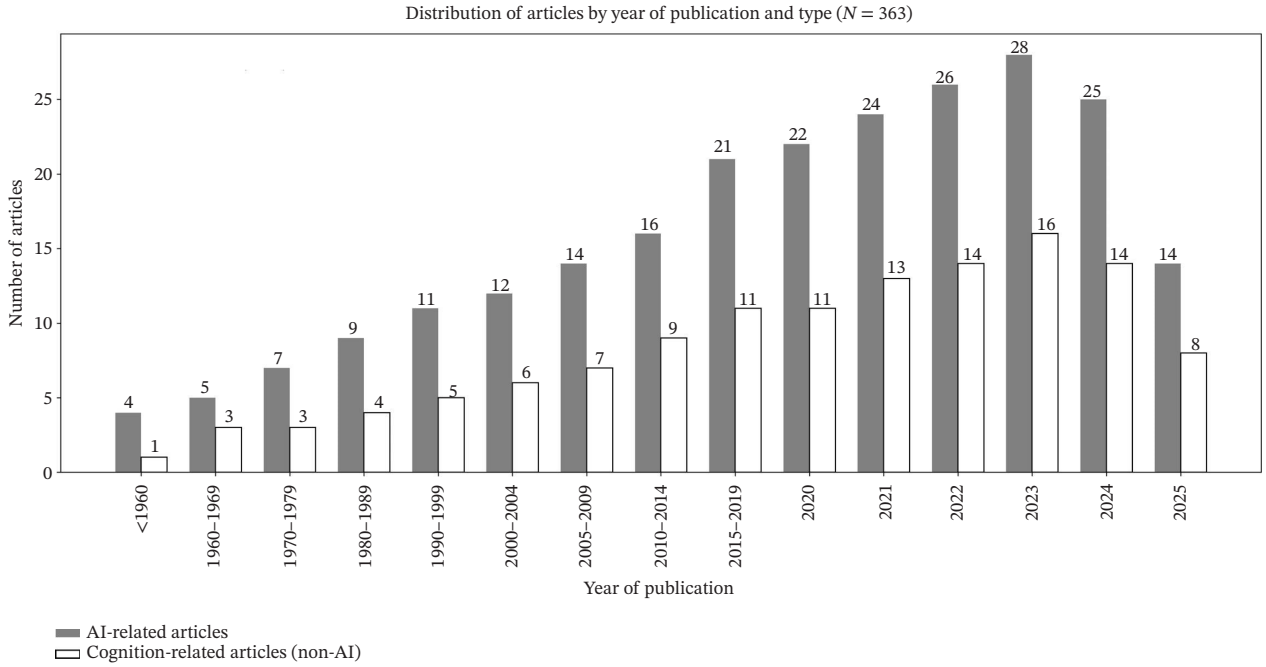

FIGURE 2 | Temporal distribution of the 363 articles retained for review, directly comparing human and AI cognition, plotted by the publication year. Bars are colour-coded to distinguish AI-related papers (studies of artificial cognitive systems) from cognition-only papers (human cognition without an AI comparison). The sharp post-2019 rise coincides with breakthroughs in deep learning and LLM research. Data for 2025 include publications up to August 2025. Note: Seven additional articles and reports providing methodological support or statistics on AI usage support are not displayed in this temporal analysis.

for the resurgence of DL. This progress led to the development of convolutional networks [88] and was further revolutionised by the transformer architecture [89]. These breakthroughs shifted AI from controlled laboratory tasks to large-scale, data-driven pattern learning across diverse domains. In parallel, researchers have pursued the integration of symbolic reasoning with neural approaches through frameworks such as logic tensor networks [90] that combine first-order logical reasoning with DL.

In his seminal book, *Artificial Intelligence—A Personal View*, Marr [91] outlined an integrative perspective to understand both natural (human) and artificial cognitive systems [92]. Concurrently, McCarthy introduced formal methods for symbolic reasoning, which became a basis for AI programming [93]. Together, their efforts shaped AI and promoted human cognition as a model for intelligent systems. This interplay between cognitive science and AI has continually evolved through a bi-directional exchange where advancements in one domain often lead to progress in the other [94].

The belief that AI and human intelligence are increasingly similar is often resisted due to concerns about control [29] and our (mis)understanding of consciousness [95]. Many people prefer to see AI as fundamentally different from humans to maintain a clear distinction between technology and humanity [96, 97]. As AI continues to improve, these boundaries blur. AI is already performing at or above human levels in a wide range of cognitive tasks [98] and even in some aspects of creativity (see Hubert et al. [99]). The cognitive and even philosophical questions arising from this overlap, such as whether AI can be truly self-aware or capable of causal reasoning, remain debatable. Several key points of convergence exist between AI and human intelligence or between artificial and human cognition [80, 100] (see also Simon [75]).

The information-processing framework conceptualises cognition as a sequence of operations that transform inputs into meaningful outputs through various mental functions [76] (see also Neisser [72]). Within this framework, core cognitive processes can be distinguished and serve as the analytical lens for comparing human and artificial cognition. Learning mechanisms involve the ability to acquire new knowledge, abstract patterns from experience and adapt behaviour through reinforcement and feedback (see, e.g., Gilchrist [101]; Tenenbaum et al. [102]). Pattern recognition and perception refer to the processes by which individuals detect structure in sensory input, enabling the interpretation of visual, auditory and other modalities of experience (e.g., Findlay and Gilchrist [103] for visual perception). Cognitive load and multitasking refer to the constraints on information-processing capacity and the mechanisms for allocating limited resources across concurrent demands (e.g., Cowan [104]; Lavie [105]; Wickens [106]; Anderson [107]; Hommel et al. [108]). Memory processes encompass the mechanisms through which information is encoded into long-term storage, maintained and later retrieved or lost over time, with retrieval often being reconstructive rather than exact (e.g., Schacter and Addis [109]; Tulving [110]). Finally, reasoning and decision-making include the ability to draw inferences, make predictions and select among alternatives, often under uncertainty and within cognitive constraints [111]. By examining these cognitive processes across human and artificial systems, the information-processing framework enables researchers to identify the shared mechanisms of intelligent behaviour and differences in how biological and computational systems achieve similar

TABLE 1 | Comparison between artificial and human cognition.

Cognitive factor	Human	Artificial	Key insights
Learning mechanisms	Learns from experience; efficient generalisation from few examples; metacognition aids reflection.	Learns from large datasets; uses metalearning.	Both adapt through feedback; AI is better at scale, humans in abstraction.
Pattern recognition	Uses priors and context to detect the structure; flexible in novel domains.	Uses statistical models and feature extraction; limited in generalisation.	Both extract regularities; humans integrate context, AI uncovers the latent structure.
Visual perception	Constructs layered representations; integrates sensory input with expectations and context; maintains constancy.	Encodes visual regularities; confuses reflectance illumination.	Both use layered and hierarchical processing; AI can incorporate biologically inspired adaptations; AI metameric failures remain.
Cognitive load and multitasking	Limited by working memory; relies on metacognitive control and task-switching strategies.	Handles parallel tasks; agentic AI can achieve goal-oriented reallocation.	Multitasking involves cost–benefit trade-offs in both systems. AI multitasks efficiently; humans adapt strategically despite limits.
Memory and forgetting	Memory is shaped by experience and embodiment; forgetting is gradual and reconstructive.	Distributed memory; vulnerable to interference; uses replay and consolidation.	Both systems reconstruct and are subject to interference; AI can benefit from biologically inspired mitigation strategies.
Reasoning, problem solving and decision-making	Combines data-driven probabilistic reasoning, inductive/deductive/abductive reasoning, Bayesian sampling and heuristics.	Relies on probabilistic, data-driven models but can use heuristics; fragile reasoning in novel, abstract domains.	Both systems prone to biases. Performance depends on complexity and contextual awareness.

cognitive outcomes [16]. Mapping and modelling these functions contribute to improving our understanding of human cognition and inform the design of artificial systems [112].

6.1 | Learning Mechanisms

Both humans and AI rely on learning from experience. This fundamental similarity is evident in how AI systems, especially those grounded in ML and DL, can learn and adapt over time based on the data they process [85]. Neural networks reproduce how neurons respond to stimuli, paralleling brain processes for information processing, memory formation and knowledge representation. Despite these similarities in learning patterns, significant differences remain in how information is processed and interpreted. AI systems are very efficient at extracting patterns from both structured (like databases) and unstructured data (such as text or images). However, human processing incorporates layers of subjective experience, cultural context and embodiment that extend beyond what even large datasets can achieve. For instance, AI can analyse millions of news articles and social media posts to detect sentiment. However, it may still miss cultural subtleties, which require an understanding of the underlying context and intent [113].

In the human brain, synaptic plasticity, such as long-term potentiation, underlies Hebbian learning [114, 115], and dopaminergic reward signals guide both model-based and model-free reinforcement learning. This allows humans to adapt their behaviour via feedback [116]. Humans also show so-called one-shot learning by forming concepts from minimal examples [117, 118] and engage in metacognitive processes, such as self-explanation, that deepen learning through reflection [119, 120]. In AI systems, supervised learning adjusts network parameters using labelled training data [78], while unsupervised methods extract structure without explicit labels [121, 122]. Reinforcement learning enables AI to learn from trial and error over time [123, 124]. Traditional deep neural networks (DNNs) are notably data-intensive compared to human one-shot learning. However, recent advances have begun to narrow this gap: Metalearning techniques such as model-agnostic metalearning (MAML) [125] and memory-augmented architectures [126] enable AI to learn from fewer examples. Furthermore, transformer-based LLMs [89] exhibit few-shot learning through in-context learning (ICL), where they adapt to new tasks from examples provided in the prompt without updating model parameters [127].

However, differences in learning efficiency between humans and AI may be partly attributable to timescale rather than mechanism alone. The heavy reliance of standard DNN on large datasets represents a practical limitation due to compressed timeframes in which AI is expected to learn rather than a cognitive flaw [128]. Humans also accumulate a vast amount of experiential 'data' throughout development and gradually refine their ability to generalise with less new input. Humans integrate compositional, hierarchical models to generalise from sparse examples [79, 102]. The learning efficiency gap likely reflects a combination of temporal compression and architectural differences.

Limitations in AI learning can be mitigated by integrating cognitive principles. Sense et al. [9] demonstrate that incorporating human-inspired learning mechanisms (such as memory decay functions and forgetting curves) improves the ability to make accurate predictions with limited data. For instance, they developed a model that simulates human-like spacing effects in memory retention, which shows superior performance on tasks with sparse training examples [129]. As mentioned, AI models trained with context-awareness through techniques such as metalearning replicate how humans integrate knowledge into new learning experiences [130–132]. AI learning mechanisms can also benefit from internal cognitive processes similar to those observed in human cognition. Lombrozo [120] describes this process as learning by thinking, in which knowledge emerges from self-explanation, mental simulation and analogical reasoning rather than direct observation. AI systems can exhibit similar behaviour, as seen in chain-of-thought prompting, where models improve performance by generating intermediate reasoning steps without additional data [133, 134]. These introspective loops echo human self-reflection by testing whether internal representations maintain coherence with the inputs that generated them.

6.2 | Pattern Recognition

The mechanisms underlying learning are intimately connected to pattern-recognition capabilities. While learning involves acquiring and adapting knowledge through experience, pattern recognition represents the application of that learnt knowledge to identify regularities and make predictions. Both humans and AI systems rely on statistical regularities extracted during learning to inform pattern detection. Although humans have traditionally been credited with superior pattern-recognition abilities, evidence from cognitive science suggests this advantage may be overstated [135]. Human pattern recognition draws on the probabilistic processing of prior experiences. However, it incorporates additional mechanisms, including compositional reasoning and model-based inference (see also Kidd et al. [136]; Tenenbaum et al. [102]). According to Sanborn and Chater [137], humans generate plausible hypotheses through approximation rather than exhaustive probabilistic computation [138].

AI systems employ explicit statistical techniques and computational power to detect correlations within datasets. Similarly, humans process patterns through neural mechanisms that are equally shaped by statistical regularities in our environment [139]. AI pattern recognition relies primarily on learnt statistical correlations and hierarchical feature extraction. The human cognitive system integrates multiple sensory inputs with prior knowledge and contextual understanding through predictive processing [140], where the brain continuously generates and updates predictions based on incoming sensory information [141]. This predictive framework allows humans to recognise patterns across modalities and in ambiguous situations where context provides critical cues.

The evolution of pattern-recognition capabilities in AI systems provides insights into cognition. DL systems can detect subtle statistical regularities that might escape human perception [142]. Convolutional neural networks [88] were the first to match and surpass human accuracy in object recognition tasks. For instance, in medical imaging, AI systems can identify patterns indicative of disease at a level that may exceed human expert performance [143]. However, it may be difficult for these AI systems to identify patterns in novel situations that deviate significantly from their training data. Conversely, humans show substantial flexibility in adapting pattern-recognition strategies

to new domains. More recent AI techniques address this limitation by transferring knowledge learnt in one context to new situations [144]. For example, models that learn relationships between images and text can recognise new concepts without specific training on those concepts by matching visual features with textual descriptions [145].

6.3 | Visual Perception

Visual perception provides a concrete example of human and AI pattern recognition in action. The ability of the visual system to extract meaningful information from sensory input exemplifies how pattern recognition operates under real-world constraints of noise, ambiguity and contextual variation. By examining visual processing specifically, we can observe how the general principles of pattern recognition manifest in a concrete perceptual domain where both humans and AI systems have been extensively studied and compared.

In human vision, the ability to detect and maintain consistent representations of objects and surfaces across varying conditions (e.g., shadows, occlusions or shifts in viewpoint) is supported by perceptual mechanisms that disentangle lightness, reflectance and illumination [146]. Classic theories posit a hierarchical structure in visual processing: low-level edge detection, mid-level perceptual grouping and high-level object inference [147, 148]. The human visual system operates through layered representations, integrating both bottom-up and top-down information to maintain coherence in visual scenes [74, 101]. These mechanisms reflect the capacity of the human visual system to construct multilayered representations [149] and to use contextual anchoring (e.g., Gilchrist and Soranzo [150]), thereby allowing stable interpretations even when sensory input is incomplete or ambiguous (see Ullman [151] for work on perceptual organisation).

Contemporary AI systems, particularly DNNs for computer vision, often reach high levels of accuracy in controlled tasks. While human perception relies primarily on global shape, many deep networks depend more on local texture cues. Their representations tend to be fragile when faced with image perturbations, unfamiliar viewpoints or contextual shifts. These conditions typically pose no problem for human observers [152]. Developments in computer vision have begun to address these limitations by incorporating biological principles [153]. For example, models that include temporal adaptation and recurrent processing show improved robustness to dynamic visual conditions and can better track objects under occlusion. These architectures introduce features inspired by the human visual system, such as the ability to update internal representations over time in response to new input. This temporal integration allows the system to reproduce the gradual accumulation of perceptual evidence.

Agrawal et al. [154] have shown that incorporating human fovea-inspired sampling improves performance on fine-grained visual recognition. Their model mimics the fovea process by concentrating high-resolution processing on the centre of the image and using lower resolution in the surrounding areas. Similarly, Lin et al. [155] developed models that separate images into broad and detailed regions and then use diffusion-based methods to enhance brightness, contrast and sharpness [156]. Such approaches explicitly replicate human perceptual constancy mechanisms, improving the perceptual coherence and generalisation of AI vision systems. Despite their impressive performance, deep neural models still fail to fully capture human perception. Humans disentangle reflectance, illumination and brightness to perceive stable surface properties across lighting conditions [157], whereas AI models tend to conflate these components and process them as overlapping visual features [158, 159]. Even though visual perception in both humans and AI involves layered and adaptive pattern recognition, these findings show that achieving human-like internal representations remains a challenge for artificial vision systems.

6.4 | Cognitive Load and Multitasking

Beyond learning and pattern recognition, cognitive systems must also manage multiple tasks and inputs simultaneously. Cognitive load refers to the mental effort required to manage, process and prioritise task-related information. Lavie [105] refines this understanding by showing that perceptual and cognitive control interact to determine attentional resource allocation and distraction susceptibility. Salvucci and Taatgen [160] proposed that multitasking performance is constrained by the availability and allocation of resources across competing task demands. Managing multiple tasks simultaneously places significant demands on cognitive resources, often resulting in diminished efficiency and increased errors.

Human multitasking involves metacognitive elements: continuous self-reflection, assessment and adjustment of task priorities based on internal performance monitoring. This metacognitive capacity enables humans to dynamically recognise limitations and adjust their resource allocation [160]. However, humans are limited in multitasking as they attempt to coordinate and prioritise competing sensory, cognitive and motor demands. For instance, driving while engaging in a phone conversation requires continuous recalibration of priorities across visual, auditory and motor systems. Such tasks highlight how humans can adaptively manage overlapping information streams [161] despite resource constraints [162].

AI systems process parallel information streams differently by leveraging distributed computational resources without experiencing the same resource bottlenecks (e.g., Thakur et al. [163]). Although AI systems handle multitasking through parallel processing, they face challenges in dynamic task prioritisation and resource allocation in unstructured environments, a process akin to the so-called executive functions that characterise human multitasking despite our limitations (see Hodgetts et al. [164] for a discussion). Both humans and AI systems experience slower processing constraints under high information load; however, the nature of these constraints may differ. Human performance degrades in both speed and accuracy under multitasking situations. This decline is well-documented in neuroergonomic and cognitive psychology research, which shows that increased demands significantly impair human task performance [165]. In contrast, AI systems encounter computational bottlenecks such as memory constraints or throughput saturation [166, 167]. This suggests an asymmetry in how humans and machines respond to information overload and multitasking.

The future of AI multitasking research points towards agentic AI systems capable of autonomous task management and priority assessment that mirror human executive functions and dynamic

cognition [168]. Some adaptive interfaces and task-switching algorithms can reduce cognitive load by dynamically reallocating resources [169, 170]. By examining how humans and AI systems manage cognitive load, researchers can develop more effective collaborative systems that leverage the complementary strengths of both human flexibility and computational processing power (see [171] for an example in human–robot interaction).

6.5 | Memory and Forgetting

Human and AI memory systems exhibit several similarities in how they operate and fail. Both systems rely on distributed architectures to store and retrieve information, with DNN explicitly inspired by the human memory structure [87, 172]. In distributed architectures, information exists across multiple connected nodes rather than in a single location; memories in the brain are encoded across networks of neurons [115, 173]. Neural networks use interconnected layers to encode and process information with learning driven by adjustments to these connections [142], much like the brain stores memories through interconnected neural pathways [174].

One significant functional similarity between human and AI memory systems lies in the role of interference. In humans, proactive interference occurs when older memories disrupt the acquisition of new knowledge, whereas retroactive interference happens when new learning hinders the retrieval of older information [175]. These processes result in the gradual weakening of memory accessibility rather than complete deletion [176]. Similarly, in AI, new learning can overwrite previously acquired knowledge. This phenomenon, known as 'catastrophic forgetting', has posed a major challenge for neural networks. For example, a neural network trained to recognise cats and then retrained on dogs might completely lose its ability to identify cats. Unlike the gradual forgetting in humans, early AI systems would experience abrupt memory loss when tasks overlapped [177].

Advances in the field have improved the ability of AI systems to manage interference better. For instance, an AI system using meta-reinforcement learning might quickly adapt its chess-playing strategy when facing a novel opening move without forgetting its core understanding of the game, similar to a human player [178]. Tadros et al. [179] provided another example of how mimicking human behaviour may solve some AI challenges with memory and interference by demonstrating the effectiveness of replay mechanisms modelled on biological sleep. These mechanisms reactivate and consolidate older knowledge during offline phases, thereby reducing interference and supporting retention (see also Refs. [180, 181] for techniques that provide resistance against forgetting by interference).

The comparison between human and artificial systems also distinguishes between episodic and semantic memory. While episodic memory in humans involves encoding specific, contextualised experiences, semantic memory supports the abstraction of general knowledge across various contexts [110]. Most LLMs resemble semantic systems: They lack temporal or spatial anchoring but do very well at extracting statistical regularities from vast textual input. This bias towards semantic generalisation enables fluent performance but contributes to their vulnerability to so-called hallucinations (i.e., plausible but incorrect outputs that may emerge when semantic patterns are overextended).

The parallels between human and AI memory failures extend to false memories. For instance, AI hallucinations are similar to human memory errors. Humans regularly misremember events owing to suggestibility or interference from similar experiences [182]. A classic example is how eyewitness testimony can be altered by the way questioning is formulated. In the seminal work by Loftus and Palmer [183], participants were shown a video of a car accident and estimated different speeds when asked about cars 'hitting' versus 'smashing into' each other. People confidently recall entire events that never happened when these false memories align with existing knowledge structures. AI systems show similar vulnerabilities. GenAI, for example, may confidently produce plausible but incorrect responses, such as citing nonexistent research papers with believable titles and author names. These errors stem from overextended semantic networks in LLMs, noise in training data, adversarial manipulation or corruption during learning (e.g., mislabelled data) that leads to false associations. The phenomenon represents not a flaw specific to artificial systems but rather a major challenge of any associative memory system.

Memory systems face key challenges in prioritising and consolidating information. These systems operate as a dynamic filtering system [184] guided by relevance and utility [185] as well as emotional salience [186]. Empirical evidence suggests that human working memory implicitly tunes forgetting rates based on information relevance [187]. Forgetting is not merely a limitation; it enables cognitive flexibility, helps filter irrelevant information and prevents overload [188, 189]. Cowan et al. [190] describe consolidation as a mechanism where significant experiences receive preferential processing for long-term storage. Emotionally powerful events like traumatic experiences are more readily consolidated into lasting memories than routine activities [191]. Similarly, reward-based prioritisation enhances the retention of motivationally relevant information [192]. Students remember material better when taught its future relevance, and people recall objects associated with rewards more than neutral objects [193].

AI systems also exhibit forms of selectivity and consolidation, although approaches vary. Traditional symbolic AI systems rely on explicit, rule-based mechanisms for prioritisation. In contrast, subsymbolic approaches such as neural networks can develop implicit prioritisation through training. Although AI systems, both symbolic and subsymbolic approaches, can be designed to prioritise high-probability patterns in data and de-emphasise rarer information, they typically operate with less contextual flexibility than human memory [194]. This limitation can lead to overfitting, where, for instance, neural networks become too specific to their training data and fail to generalise. Richards and Frankland [195] describe a similar trade-off in human memory between specificity and generalisation. Unlike humans, who integrate emotional salience, utility and relevance when encoding and remembering information, symbolic AI systems require explicit programming for these features, while neural networks can learn them through training. Memory systems must balance retention with adaptability and context sensitivity. AI systems are beginning to achieve similar flexibility through architectural and training innovations [196]. Understanding both human and machine approaches to memory can help us to design systems that combine their respective strengths and mitigate their limitations. For instance, AI can benefit from

implementing variable forgetting rates based on information utility rather than processing all data equally [197] (see, e.g., Lavoie-Hudon et al. [198]).

6.6 | Reasoning, Problem Solving and Decision-Making

Despite being often used interchangeably, problem solving and decision-making represent distinct yet closely connected cognitive processes. Decision-making involves evaluating and selecting, under uncertainty, among alternatives. These alternatives can be relatively static and independent choices [199–201] or dynamic and interconnected options [202]. Decision-making can be broadly categorised into three interrelated types, each presenting distinct challenges for both humans and AI, such as classification decisions, judgement decisions and dynamic decision-making (which, in turn, can be divided into fast, time-pressured decision-making and complex decisions based on elaborate reasoning).

Problem solving involves identifying discrepancies between current and desired states and then generating strategies to bridge this gap [203]. A relevant framework for understanding problem-solving processes is the distinction between complicated and complex problems [204, 205]. Complicated problems contain many components but follow predictable patterns and are governed by fixed rules. However, complex problems involve numerous interconnected variables, unpredictability, interdependencies and multiple potentially conflicting goals [206].

As a cognitive process, reasoning involves drawing inferences from available information. In cognitive science, reasoning is often categorised into three types: deductive (applying general rules to specific instances), inductive (deriving general principles from specific observations) and abductive (forming the most likely explanation from incomplete information). These reasoning processes underpin problem-solving and decision-making activities [111]. These processes are also affected by constraints such as time pressure, uncertainty and cognitive load, which influence whether individuals rely on deliberate reasoning or fast, heuristic-based thinking [207].

Classical AI research addressed problem solving and decision-making via distinct approaches, each with different architectural foundations and application domains. Early symbolic AI, such as Newell and Simon's General Problem Solver [82], focused on general-purpose problem solving by systematically exploring possible solutions and progressively reducing the gap between the current state and the desired goal. These systems targeted well-defined domains such as mathematical puzzles and strategic games by representing problems as structured sequences of steps with explicit rules for moving from one state to another [73]. Later AI research shifted towards domain-specific decision support through rule-based expert systems. These expert systems provide support to human decision-making by encoding domain knowledge from human experts into explicit rules to recommend actions under uncertainty.

AI has reached and even surpassed human performance on a wide range of cognitive tasks that require advanced thinking and reasoning. Chess, once the gold standard for measuring AI capability, is now dominated by AI [208]. AI systems master strategic thinking in games like Go, making moves that even human experts could not predict [143] and succeed even in games like Diplomacy, which involves negotiation [209]. Similarly, the AI system Pluribus defeated professional poker players in six-player no-limit Texas Hold'em, which demonstrates the ability to make decisions under uncertainty [210]. Another AI system, AlphaStar, reached the grandmaster level in the game StarCraft II, outperforming 99.8% of human players in this real-time strategy environment [211]. Hagendorff et al. [212] have shown that advanced LLMs can even outperform humans on cognitive reflection tests designed to elicit human reasoning errors [213]. In professional domains, AI models have achieved passing-level performance or high scores on licencing-style and qualification assessments, such as the Uniform Bar Examination, where GPT-4 scored well above passing thresholds across all components [214].

These AI accomplishments largely reflect performance in structured problem spaces, where success depends on fast pattern recognition and data synthesis. These successes often occur in constrained environments with clear rules, not in open-ended, ill-structured domains requiring contextual adaptation. Although models like GPT-4 show impressive results on reasoning benchmarks, some researchers argue that their inferences rely on pattern matching across training sets rather than true causal or analogical reasoning [215], although this interpretation remains a matter of debate. The distinction between complicated and complex problems may partly explain why AI systems show superior performance in chess, Go and professional exams yet face difficulties with open-ended scenarios requiring contextual judgement. Nevertheless, as mentioned, AlphaStar, a deep reinforcement learning system, can handle dynamic and relatively complex decision-making, as demonstrated by its performance in StarCraft II. It is important to note that LLMs were not explicitly designed for problem solving or decision-making. While they demonstrate capabilities on various benchmarks, such performance stems from emergent properties of language modelling rather than architectures built for reasoning [26, 27, 215].

One key distinction in how AI and human systems approach problem solving is that AI has the computational capacity to deal with very large datasets. In contrast, due to their limited capacity, humans are likely to restructure problem spaces to find efficient, though suboptimal, solutions. This is in line with Simon's 1955 concept of satisficing, where decision-makers produce 'good enough' solutions rather than optimal ones, given constraints in time and cognitive resources [216] (see also Goodrich et al. [217]). Even AI systems, despite their ability to process large datasets, face practical real-world constraints (such as processing time, energy costs, response latency) that make computational efficiency valuable. Hélie and Pizlo [218] suggest that AI can benefit from incorporating human-like heuristics to improve real-time decision-making (see also Gigerenzer [219]). Research on models that account for cognitive and computational costs shows that simple heuristics can achieve near-optimal outcomes when operating under constraints such as limited time, memory or processing capacity [220, 221]. This body of research suggests that computationally simple heuristics may sometimes outperform complex optimisation algorithms in conditions characterised by uncertainty and limited resources [222, 223].

However, heuristics can also introduce biases that are potentially detrimental to reasoning [224]. Evidence suggests that AI

systems, like humans, are subject to cognitive biases that can impact problem solving and decision-making [225]. Research indicates that LLMs can reproduce human-like errors when making judgements based on anchoring, availability or confirmation cues [226]. For instance, Nguyen found that LLMs, including GPT-4 and Claude, exhibit anchoring bias, that is, the tendency to rely heavily on initial information when making judgements. When asked to make financial forecasts, these AI systems produced significantly different predictions based on whether they were first exposed to high or low numerical values, which is a well-documented bias in human decision-making. Attempts to mitigate this bias through explicit instructions or structured reasoning prompts showed limited success. These parallels suggest that the tendency towards cognitive biases may arise not only from biological constraints but also from the architecture of learning systems more broadly.

7 | Reframing Expectations: Misconceptions About AI and Human Cognition

Public opinion of AI oscillates between overestimating its capabilities and underestimating its possibilities, which results in both inflated expectations and dismissive criticisms. To address misconceptions, we organise our analysis around three interconnected themes. First, we examine whether AI systems truly understand language and concepts or merely identify statistical patterns. We re-examine the 'stochastic parrot' critique and explore the nature of logical reasoning in both human and artificial systems. Second, we review evidence of shared cognitive vulnerabilities. We show that both humans and AI make errors and exhibit biases across multiple cognitive processes, albeit with different manifestations and consequences. Third, we investigate the questions of transparency and agency, which include opacity, free will and creativity. Neither human nor artificial systems operate with the complete transparency or independence often assumed. Table 2 provides a summary of the key comparisons reviewed in this section.

7.1 | Is Generative AI Merely a 'Stochastic Parrot'?

Computer scientist Emily Bender (see Bender et al. [10]) suggests that GenAI produces statistically likely outputs based on its training data rather than understanding or reasoning. According to this view, AI does not understand anything; it can only mimic human language superficially without any comprehension or insight. However, this critique is perhaps overly reductive. The key question is whether statistical pattern matching necessarily precludes understanding, or whether understanding itself might emerge from pattern-based processing. GenAI systems do not merely repeat phrases they have encountered; they synthesise new responses by recombining large amounts of information in novel ways. The ability to generalise from data and generate appropriate responses to unseen prompts suggests a level of abstraction that exceeds simple parroting. Humans can typically produce recursive syntactic constructions (such as nested relative clauses) without explicit awareness of the underlying grammatical rules [227]. Similarly, GenAI systems can generate recursive linguistic structures by relying on statistical regularities acquired through training. One can argue that, in both cases, recursion emerges from implicit, pattern-based processing rather than from explicit symbolic computation [228].

GenAI systems exhibit several capabilities that challenge the 'stochastic parrot' characterisation. Contextual understanding can be ascribed to their ability to maintain awareness across conversations, adapting responses based on dialogue history and user intent, enabling them to follow conversational logic and provide contextually appropriate answers. Beyond this, their analytical output capabilities allow them to generate solutions, analyse various forms of information and provide insights, from creating stories to generating and debugging code. Although these abilities are grounded in pattern recognition, they showcase the production of new and useful outputs. Furthermore, larger and more complex models exhibit emergent capabilities that were not explicitly programmed, such as performing arithmetic or following logical steps, hinting at forms of reasoning that developed organically from model training and data exposure rather than direct coding [214].

Critiques of AI as a 'stochastic parrot', such as those by Thierry [229], typically rely on the argument that these systems lack true understanding or intent. Thierry argues that AI merely 'masquerades as minds' and generates superficially coherent language without awareness, meaning or goals. These critiques may serve as a safeguard against uncritical anthropomorphism. However, they often rest on an implicitly dualistic conception of human cognition: one in which thought is guided by an internal, quasispiritual core capable of transcending statistical associations [230]. This perspective resonates with long-standing efforts in cognitive science to 'banish the homunculus', that is, the metaphorical inner agent invoked to explain intention, control or understanding. As Verbruggen et al. [231] argue, appeals to internal controllers in models of behaviour often mask rather than solve theoretical problems; they can be replaced by more mechanistic, distributed accounts of action control. Similarly, Hazy et al. [232] proposed a biologically plausible framework for working memory and cognitive control that operates through interactive neural dynamics within prefrontal and basal ganglia circuits. These models show that cognitive functions such as memory and decision-making can be achieved without invoking a central executive controller. This challenges critiques that position human intelligence as categorically distinct from the distributed inference systems of modern AI. These arguments echo Margolis' [233] philosophical rejection of explanatory regress in so-called mentalist models and reinforce that neither biological intelligence nor AI depends on a metaphorical homunculus to be meaningful.

Theories of the 'Bayesian brain' (e.g., Knill and Pouget [12]) further challenge the assumption that human understanding is categorically distinct from statistical inference. In this framework, the human mind operates as a predictive inference engine that constantly generates and updates hypotheses about the world based on sensory feedback. Chater and Oaksford [11] similarly describe human cognition as a form of Bayesian sampling that relies on heuristics and approximations rather than optimal logical solutions. Even highly specialised human abilities, such as face recognition, illustrate this pattern-matching tendency. Thierry et al. [234] showed that neural responses associated with face perception are not exclusive to human faces; similar activation patterns occur when people view cars, door handles or other stimuli that share configural properties with faces. These findings support the argument that

TABLE 2 | Key comparisons related to human and artificial cognition.

Theme	Human	Artificial	Key insights
Is generative AI merely a 'stochastic parrot'?	Uses pattern-based language processing, often without awareness of underlying grammatical rules. Understanding may emerge from statistical regularities.	Generates outputs through statistical associations and pattern recombination; exhibits some contextual awareness and emergent capabilities.	Both rely on pattern-based processing. Understanding might emerge from statistical mechanisms in both systems.
Is AI capable of deductive logic?	Often deviates from formal logic. Prefers probabilistic and inductive reasoning over deduction. Integrates deductive, inductive and abductive reasoning flexibly but with logical inaccuracies.	Achieves deductive-like reasoning via probabilistic pattern-matching. Solve some classic syllogisms by leveraging linguistic patterns that reflect logical structures. Often fails with unfamiliar formats.	Neither reasons through pure formal logic. Both achieve functional reasoning. Performance degrades with unfamiliar formats in both systems.
Are humans rational, fully self-aware and unbiased?	Metacognition is limited; often unaware of biases and limits.	Biases from training data and algorithmic design choices. Self-monitoring is absent unless designed.	Both systems exhibit bounded rationality and metacognitive limitations.
If to err is human, what does it mean that AI errs too?	Prone to errors, biases and perceptual illusions arising from predictive, heuristic-driven processing.	AI systems are vulnerable to errors due to data limitation, adversarial inputs or overfitting.	Human reactions (algorithm aversion/appreciation) shape AI adoption. AI errors pose systemic risks because of replication across instances.
Is free will uniquely human, or is it an illusion?	Agency is often a post hoc narrative. Decisions are influenced to some extent by genetics and environment.	AI operates without subjective experience or volition. Behaviours reflect learnt patterns, not intentional choice.	Free will and consciousness may lie on a continuum rather than a binary divide. Emerging perspectives on artificial consciousness.
Is AI capable of creativity?	Goal-driven, emotionally informed, socially grounded.	Recombines learnt patterns; can exhibit divergent thinking.	Both recombine prior knowledge; humans add intent and emotional salience.
Are human decisions fully transparent?	Humans often rely on opaque heuristics. Explanations of decisions are often not fully accessible and thus postrationalised.	Many complex models are neither interpretable nor explainable. Full transparency often remains elusive.	Transparency is limited in both. The complexity-comprehension paradox limits the effectiveness of full AI transparency.

human brains, like AI systems, may operate as adaptive pattern recognisers.

Dennett [13, 14] has argued that cognition does not require an internal self that 'understands' in a conscious, deliberative sense; rather, intelligence and agency emerge from the interaction of simple components under appropriate constraints. Baars [235] similarly critiques the persistent mind–body dualism in interpretations of cognitive processes and calls for a more integrated, mechanistic account of the mind [236]. Current AI systems do not exhibit embodied experience, self-reflective consciousness, intrinsic motivation and autonomous goal formation. However, drawing a strict line between human and artificial cognition based on criteria themselves undermined by findings in neuroscience and psychology represents a missed opportunity for a comparative understanding of biological and artificial minds. Seth [237] argues that the true artificial consciousness is unlikely under computational AI paradigms. Consciousness, they contend, stems from our biological nature, although it becomes more plausible only if machines grow increasingly brain-like or life-like. Such a shift would also bring profound ethical challenges (see also Mahowald et al. [26]).

7.2 | Is AI Capable of Deductive Logic?

Felin and Holweg [238] argue that human reasoning is uniquely forward-looking, fundamentally theory-driven and causal and built for abstraction, counterfactual reasoning and deep sense-making [239, 240]. They suggest that AI, being data-driven (unless knowledge-based rules are coded and implemented by humans) and statistical, cannot replicate human-like reasoning [59]. However, this position is challenged by Chater [241], who argues that human cognition itself operates through shallow statistical processing similar to modern AI systems. He also argues that human cognition operates reactively and constructs rationalisations. This view is in line with the concept of bounded rationality [216, 242], which suggests that human reasoning is constrained by cognitive limitations, incomplete information and time pressure, leading individuals to rely on heuristics rather than purely rational computation [243, 244]. According to the Bayesian sampler hypothesis, humans do not store complete probability distributions but instead sample from past experiences to approximate probabilistic reasoning [245].

Deductive reasoning is the process of deriving specific conclusions from general premises through formal logical rules and has long been central to logical thinking [246]. From Aristotelian syllogisms to contemporary computational systems, deductive reasoning remains essential to developing structured arguments. Syllogisms illustrate this approach through their characteristic two-premise structure that leads to a conclusion, with specific logical forms (e.g., 'all A are B', 'some A are B') providing a systematic framework for analysis [247]. Yet human deductive reasoning frequently deviates from formal logic due to cognitive biases and heuristics. For instance, when presented with 'no police dogs are vicious; some highly trained dogs are vicious; therefore, some highly trained dogs are not police dogs', most people reject this valid conclusion because it conflicts with their beliefs about police dogs [248]. Confirmation bias adds to these errors, as people tend to seek evidence that supports their preferred conclusions [249].

Oaksford and Chater [250] suggest that humans often prefer inductive reasoning and probabilistic approaches when dealing with uncertain information rather than strictly adhering to deductive logic (see also Holland [251]). This shift towards probabilistic reasoning reflects a natural inclination to deal with uncertainty, which aligns with the Bayesian framework of human cognition [252]. Many real-world problems require a combination of deductive, inductive and abductive reasoning rather than pure deduction alone. Humans typically integrate contextual cues, prior beliefs and background knowledge in their reasoning processes by generating general rules from specific observations (inductive), applying general principles to specific cases (deductive) and forming the best explanations from incomplete data (abductive). This integration makes human reasoning flexible but also prone to logical inaccuracies [253]. AI systems can perform reasonably well across this spectrum of reasoning by generalising from patterns observed in large datasets.

GenAI is assumed to be unable to perform the rule-based reasoning that underpins deductive logic. According to some researchers, because GenAI depends on probabilistic patterns instead of explicitly applying deductive rules [254], it cannot fully engage in formal logical reasoning (see Ref. [255]). However, a number of documented outcomes of GenAI seem to challenge this notion. For example, LLMs like GPT-4 have shown emergent deductive capabilities by leveraging their extensive training on linguistic data. Classic syllogisms can be resolved by GenAI, not through formal logical computation but rather from patterns derived from linguistic data that inherently reflect logical structures. Even though the underlying mechanism is probabilistic (relying on the likelihood of words and phrases co-occurring), the result is often logically sound because language and reasoning are tightly intertwined [227, 256]. This approach allows GenAI to handle tasks that involve deductive-like processes, especially in familiar, well-documented domains [257]. Nevertheless, as Cheng et al. [258] observe, LLMs are unreliable when constructing unseen logical structures. In a series of experiments, Jiang et al. [259] reveal limitations in how LLMs process logical problems. For instance, when presented with the classic 'Linda problem' (a well-known conjunction fallacy test), GPT-4 correctly identified that the probability of Linda being a bank teller must be higher than her being both a bank teller and a feminist. However, when researchers changed 'Linda' to 'Bob' while preserving an identical logical structure, performance dropped significantly (see also [260]).

Despite these shortcomings in current LLMs, emerging innovations in AI architecture design offer pathways for improving logical reasoning capabilities. By explicitly integrating logical rule-based structures into neural network frameworks, researchers are developing approaches that may overcome the pattern-matching limitations of current systems [261]. For instance, the dual-agent framework proposed by Du et al. [262] shows the potential for LLMs to refine their deductive reasoning capabilities. This setup involves collaborative agents—one acting as a questioner and the other as an answerer—working together to navigate logical tasks like the '20-question' game (see also Liang et al. [263]).

7.3 | Are Humans Rational, Fully Self-Aware and Unbiased?

Human understanding of the rationale behind our own decisions is often far more limited and ambiguous than we might like to admit. The idea that humans possess perfect self-awareness and

rationality is largely a myth [264]. Human metacognition, the ability to monitor and evaluate one's own thought and cognitive processes, reveals miscalibrations when subjected to empirical investigation (e.g., Rousseau et al. [265]). Empirical work has documented substantial gaps between what people believe about their cognitive abilities and their actual performance [266].

As previously discussed, psychological research has consistently shown that people often make decisions based on intuition or unconscious biases and only then try to rationalise them with logical explanations (e.g., Dennett [267]; Haidt [268]). There is ample evidence that individuals who do not possess expert knowledge and skills in a particular domain overestimate their abilities [269, 270]. Moreover, the concept of choice blindness suggests that decision-making is not always guided by stable, deeply held preferences but is instead malleable and context-dependent [138]. This metacognitive unreliability extends beyond inaccurate self-assessment of knowledge and skills to the very processes of decision-making and preference formation. Research on constructed preferences [271] and unconscious influences on choice and forecasting [272, 273] collectively suggests that individuals are frequently unaware of the factors shaping their decisions. Even though people make decisions based on complex cognitive processes, they may fail to articulate their reasoning for that decision [274]. Although humans may be more self-aware than AI can be, they often cannot fully explain their decisions.

The capacity to self-reflect, that is, to monitor, evaluate and revise one's own mental processes, is deemed by many researchers to be central to learning, decision-making and self-regulation [275, 276]. Humans engage in reflective processes to assess confidence and error as well as to revise their goals, question assumptions and gain insight, processes linked to notions of autonomy and responsibility. AI systems are beginning to exhibit elementary forms of metacognition, such as revising their own responses to improve coherence [277]. Such models can evaluate their outputs and then regenerate improved answers. Related research on system introspection explores how a model can generate a textual representation of an input (e.g., an image caption) and uses that description to regenerate the original stimulus; high fidelity between the two suggests an internally coherent representation rather than superficial pattern matching. Early instantiations of this caption-regeneration loop appear in agentic AI [278] and in techniques that train models to predict their own responses [277] as a form of computational metacognition. These developments raise important questions about whether machines can possess reflective cognition without consciousness.

7.4 | If to Err Is Human, What Does It Mean That AI Errs Too?

Humans are susceptible to visual, auditory and cognitive illusions whereby our perception is systematically misled by specific patterns or contextual cues [279, 280]. These perceptual distortions reveal limitations in human sensory processing and suggest that our cognitive system interprets information based on predictive models, expectations and prior knowledge rather than simply registering direct sensory data [281]. Visual illusions like the Müller-Lyer or Ebbinghaus effects expose how our visual system applies unconscious inferences to make sense of ambiguous stimuli [282], while auditory illusions such as the Shepard tone and phonemic restoration effect similarly highlight how our auditory processing tends to (re)construct coherent perceptions from incomplete information [283, 284]. These perceptual anomalies highlight that human perception is not an exact recording of reality but rather an active, interpretive process shaped by cognitive and neural constraints [141]. However, these errors often serve a functional role in human cognition, acting as feedback for the learning and recalibration of mental models [285].

AI systems are also vulnerable to inputs that can mislead their classification and inference process [286]. For example, small, often unnoticeable changes to an image can cause artificial vision to misidentify it entirely, a similarity to human perceptual illusions in outcome rather than mechanism [128]. AI will often make errors based on the limitations of its training data that reflect representational and statistical sensitivities. This is similar to how human decisions are influenced by past experiences and learnt information. If biased, incomplete or adversarially modified (or selected; so-called naturally adversarial) data are used for training, AI models and their predictions will reflect these problems. AI models can be manipulated by deliberate changes to input data that cause AI to make incorrect predictions. This is comparable to how humans can be deceived by illusions or manipulated by misinformation [287]. AI also exhibits intrinsic algorithmic limitations due to constraints defined by its underlying computational and training approaches (see Hooker [288]). Modern AI systems still show vulnerability when confronted with new operational domains or scenarios that significantly diverge from what they were initially trained on (e.g., Hendrycks et al. [289]). This phenomenon parallels human cognitive biases, where both AI and humans fail to adapt when faced with unfamiliar problems requiring substantial transfer of learning [79].

Acknowledging that both AI and humans can err (often similar in observable ways) supports the idea that error making is a feature of complex adaptive systems rather than a flaw exclusive to artificial models. Despite shared tendencies for error, human and AI errors differ in their systemic nature. Human errors tend to be clustered, state-dependent and context-sensitive, affected by factors such as fatigue, stress and task difficulty [279, 290]. These patterns are often domain-specific, such as making multiple errors in similar medical procedures, or condition-specific, such as making more errors when fatigued [291]. In contrast, AI often does not exhibit human-like clustering or metacognitive expressions of uncertainty, which can make failures appear unpredictably distributed from a human perspective rather than truly random. AI errors can also be differentiated from human errors by their uniformity and scale [279]. Empirical work further shows that AI systems tend to make highly similar mistakes to one another, whereas human errors are more heterogeneous across individuals [292]. When an AI system makes an error, that same error is replicated across every instance of the system, potentially affecting millions of decisions at once [293].

7.5 | Is Free Will Uniquely Human, or Is It an Illusion?

Some researchers suggest that human thoughts and decisions are profoundly influenced by deterministic factors (biological, social and environmental) that shape behaviours beyond conscious awareness [294, 295]. For example, genetics, cultural background and cognitive biases have been shown to drive much of our

thinking and decision-making, often without us realising their influence (e.g., Béchard et al. [296]). Empirical work on the timing of conscious intention, such as Libet's classic yet controversial experiments, has been argued to suggest that the feeling of making a decision is a post hoc narrative constructed to make sense of processes already in motion [297] (see also Matsuhashi and Hallett [298]).

Sapolsky [135] extends this view, arguing that the sense of agency is an illusion essential for social cohesion. In his view, self-awareness is less of a cognitive capacity and more of a self-generated story designed to make sense of our behaviour and social interactions [299]. From this perspective, many human behaviours and choices may be adaptive responses to environmental stimuli akin to advanced AI pattern recognition. Kotchoubey [300] adds to this perspective by framing consciousness as an emergent property of adaptive behaviour. He describes consciousness as a 'mental space' for simulating potential actions and anticipation. This anticipatory capacity could be argued to resemble advanced AI systems, which simulate potential outcomes based on learnt patterns without genuine intentionality. Brass et al. [301] refine discussions on free will by challenging overly deterministic interpretations of neural and cognitive processes. They argue that unconscious neural activity does not preclude free will. Their research posits that conscious intentions establish parameters for decision-making processes and facilitate the inhibition of undesired behavioural responses. Mudrik et al. [302] propose that free will operates along a spectrum rather than as a binary construct.

Research on artificial consciousness and the Theory of Mind (ToM) in AI extends these questions about the nature of agency and free will [303, 304]. ToM, defined as the ability to attribute mental states to oneself and others [305], can serve as a conceptual framework for analysing the capacity of AI systems to predict and interact with human behavioural patterns. Farisco et al. [306] argue that while human-like consciousness remains unattainable for AI, alternative forms of consciousness-like processing adapted to AI architectures are theoretically possible [307]. This possibility calls for a reassessment of traditional notions of autonomy and raises philosophical and ethical questions [308]. Philosophers and researchers now explore whether AI systems capable of exhibiting alternative forms of consciousness might redefine the boundaries of autonomy and agency in relation to free will (see Bayne et al. [95]).

These developments have implications for the debate about free will. If alternative forms of agency can emerge in artificial systems through learnt goals and adaptive behaviour [309, 310], this challenges the view that free will requires biological embodiment and phenomenal consciousness [311]. Moreover, if both human and artificial decisions emerge from accumulated patterns rather than unconstrained choice, natural and artificial agency may differ in degree rather than in nature [312]. This convergence suggests that free will, whether in humans or machines, may be better understood as context-dependent autonomy within learnt constraints rather than absolute freedom of choice.

7.6 | Is AI Capable of Creativity

Despite decades of research, the cognitive and neurobiological mechanisms underlying creative thought have remained largely underspecified, with traditional cognitive approaches failing to capture the spontaneous nature of creativity [313]. Creativity is often romanticised as a uniquely human capacity, yet recent developments in AI challenge this assumption. At its core, creativity involves generating novel and valuable ideas or artifacts, but the definition of what constitutes novelty remains contested. Humans generate 'new' ideas by recombining patterns from a large corpus of data, much of which comes from their experiences, education and interactions with the world (see Fauconnier and Turner [314]). According to Fauconnier and Turner, creativity emerges when we combine elements from different mental spaces or knowledge domains. Both humans and AI rely on prior knowledge and experiences (or training data, in the case of AI) to develop new ideas or solutions (see Leach [315]). The ability of AI to recombine data can produce outputs that seem novel because it identifies relationships within the data that might not be immediately obvious to humans [316].

The mechanisms that enable recombination involve both associative and combinatorial processes. At the associative level, human cognition relies on both common associations between high-frequency, closely linked concepts and remote associations, which require bridging distantly related ideas. Classic associative network models (e.g., Mednick's associative theory [317]) propose that creative and flexible cognition emerges from the ability to traverse conceptual space at varying distances [318]. AI systems show analogous behaviour: Transformer-based architectures encode dense manifolds where semantic proximity arises from statistical co-occurrence, enabling both local prediction and remote associative leaps (e.g., analogy generation, metaphor synthesis) [228]. Additionally, human thought frequently operates through combinatorial processes, where elements are recursively recombined into higher-order structures, as seen in conceptual blending [314], mental model construction and language syntax. Contemporary AI mechanisms such as chain-of-thought prompting [134] and compositional generalisation [228] emulate the aspects of combinatorial processing. These parallels suggest that abstraction and flexible recombination (once seen as uniquely human) may arise in any system that learns distributed representations under sufficient variation and scale.

Although we may feel capable of creating entirely new concepts, neuroscience research reveals that creativity involves dynamic integration between the brain networks responsible for memory and sensory processing [319]. These findings suggest that creative insights emerge from the novel combinations of existing knowledge. Groundbreaking ideas like democracy or abstract art could be seen as extensions or adaptations of existing principles, perspectives and observations rather than entirely new constructs [320]. Cultural [321] and linguistic [322] contexts shape our thinking and imagination, setting a frame for what we perceive as 'new' ideas [323]. This phenomenon is often termed an 'illusion of novelty': While our ideas feel innovative to us, they remain fundamentally tied to pre-existing culture and cognition.

The capacity to take risks is often cited as a specific human creative trait that pushes beyond established boundaries in search of the unknown. Innovation is conceived as involving risk-taking and exploration [98]. A musician might experiment with unconventional sounds that defy traditional genres. This willingness to take creative risks can lead to unexpected outcomes. However, the assertion that AI cannot take creative risks

by generating outputs based on probabilities is debatable [324]. GenAI has been shown to outperform humans in divergent thinking, a skill associated with creativity [99]. The use of techniques that mimic risk-taking and divergent thinking can produce creative outputs, often indistinguishable from or preferred over human-created art [325], but only when human evaluators are not told whether the creator is human or artificial [326]. Research also shows that AI systems can be tuned for novelty via functionalities that optimise for originality [327, 328]. However, human creativity is guided by the 'why' question and shaped by emotional and social considerations [98]. Whether creating art to evoke emotions or developing solutions to address societal issues, humans innovate with a sense of purpose that AI cannot (yet) replicate. From a cognitive perspective, it seems we approach creativity by planning, revising and reflecting. We set goals, question assumptions and change direction based on our evolving understanding and desires. AI can simulate goal-directed behaviour through prompts; however, it does not have the reflective capacity that guides human creativity [329].

7.7 | Are Human Decisions Fully Transparent?

There is a quest for transparency in AI models on which recommendations and automated actions are based. There is pressure on industries and governments to find solutions and set regulations to ensure that AI decision paths are understandable, particularly in complex fields like medical diagnostics and finance [330]. A critical challenge in AI development concerns balancing the need for high predictive accuracy with the demand for transparency and understandability [331]. Both explainable AI (XAI) and interpretable AI seek to improve transparency through different approaches. The goal of interpretable AI is to develop inherently understandable models, such as decision trees in medicine [332]. In contrast, XAI focuses on post hoc explanations for 'black box' models, such as generating heatmaps to highlight the most influential regions in an AI-assisted medical diagnosis.

The expectation for AI to be both accurate and explainable introduces a necessary trade-off. Researchers have extensively documented this 'interpretability-accuracy trade-off'. The key observation is that more interpretable models (e.g., decision trees) generally achieve lower predictive accuracy [333], while more accurate systems (e.g., DNN with millions of parameters) remain opaque [334]. AI is often held to a higher standard of transparency than human decision-makers, who regularly employ tacit knowledge and intuition without explicit explanations [335, 336]. Raz et al. [337] describe this as a transparency double standard; while AI systems receive criticism for their opacity, humans operate within institutional and social frameworks that provide post hoc justifications rather than real-time explanations. Human decision-making also suffers from what Rozenblit and Keil [338] termed 'the illusion of explanatory depth': We tend to overestimate our understanding of complex phenomena [339].

XAI promises to make complex models understandable to humans; however, as AI models grow more sophisticated, the very nature of their complexity often surpasses human capacity to grasp those models fully [331]. Liao and Vaughan [339] argue that as AI models become more complex, transparency efforts fail to enhance human comprehension. There is ample evidence that in the case of complex problems, not just complicated problems [203], AI-based decision support systems act more like a cognitive prosthesis to human cognition [340]. The calls for transparent AI create a paradox because, in many situations, AI can (and will be expected to) solve problems that are too complex for the human brain [341]. DL models produce outputs based on layers of interactions and feature abstractions that are beyond human comprehension [80].

XAI and interpretable AI approaches run into the complexity-comprehension paradox [342]. For interpretable AI, this paradox expresses the need to simplify models to make them comprehensible; however, this simplification inevitably means sacrificing the ability to capture complex relationships. A simple scoring system for loan decisions might be perfectly interpretable but misses patterns in financial behaviour that a more complex model could detect (e.g., Baesens et al. [343]). When explaining complex models, XAI methods provide simplified explanations mainly through data visualisation [344]. Even with tools like SHAP [345, 346] and LIME [347], humans can only partially understand the reasoning behind AI predictions and decisions [331, 348]. These methods might reduce complex models to surface-level explanations.

Oversimplification can give a misleading sense of transparency, where users feel they understand how AI operates without understanding the underlying depth. This can lead to overconfidence in some users and mistrust in others [349]. Humans are often biased when interpreting AI-generated explanations, especially in ambiguous situations. For instance, confirmation bias can lead individuals to selectively trust aspects of an output that match their expectations, potentially overlooking critical insights or anomalies. Whether we build simplicity from the start (interpretable AI) or create simplified explanations afterwards (XAI), we are constrained by the gap between AI capabilities and human comprehension [350, 351].

8 | The Emergence of (Artificial) Intelligence

The argument that human intelligence significantly advanced with the development of language [68] is in line with the idea that language is a critical tool for shaping cognitive abilities [352]. If we applied this perspective to AI and LLMs, it may suggest a parallel: Just as language catalysed the development of human intelligence and cognition, LLMs might act as a catalyst for further developing machine intelligence. Pinker [353] and other cognitive scientists suggest that language is not just a by-product of intelligence but a driving force behind cognitive evolution (see Lupyan [354]) and that language can be seen as a cognitive tool [355].

Language allows humans to express thoughts, share information and collaborate in ways that would not be possible without a common communication medium. However, human language acquisition is deeply tied to embodied sensorimotor experience, through which words acquire meaning in perception and action [356, 357]. Language enables abstract thinking, the formation of social structures and the accumulation of knowledge. To some extent, it also shapes how we think [358]. For example, different languages emphasise different aspects of the world (time, space, emotions) and thus exert a significant influence on how speakers perceive and reason about these concepts. Human intelligence evolved in a social context, and language was key to facilitating

cooperation, teaching and coordination within groups. If we draw an analogy to LLMs, one could argue that something akin to human cognition is emerging in GenAI because of the ability to process, generate and manipulate language on a large scale.

LLMs are trained on large datasets of human language that encompass centuries of accumulated knowledge. This massive linguistic exposure allows them to form networks of associations between concepts, which parallels how humans build and transmit understanding through language. Beyond language as a knowledge base, there seems to be an emergence of new abilities. LLMs exhibit capabilities that were not explicitly coded in a similar fashion to how language enabled humans to reason abstractly. These emergent capabilities seem to appear only after models reach certain thresholds of scale and complexity. Multimodal models that combine vision, audio and language may further support such emergent behaviours by adding perceptual grounding to textual learning (see, e.g., [359, 360]).

In both humans and AI, interaction and social communication through language are deemed critical in developing intelligence. In the case of AI, interactions with human users and other AI agents, as well as feedback loops and continuous learning, help refine its capabilities. The iterative process of refining models based on language data parallels the feedback humans receive through communication. Human intelligence also evolved in a social context, and similarly, LLMs can improve because of their ability to interact with multiple forms of communication. As these systems become embedded in education, healthcare and business environments through interfaces ranging from chat platforms to APIs, they increasingly participate in collaborative problem solving. As human intelligence thrives through collaboration facilitated by language, LLMs may enhance collective problem-solving abilities, particularly when embedded in collaborative platforms or teams (albeit multiagent or human–agent teams).

Whether AI could evolve like human intelligence is more speculative, but some parallels exist between human cognitive development through language and the growing complexity of LLMs. Language has been fundamental to human cognitive evolution, and LLMs are finding a viable pathway for AGI development (see, e.g., Xu and Poo [361]). However, limitations persist in areas such as causal reasoning and planning [30]. This perspective is complemented by LeCun [67], who argues that language processing alone is insufficient for developing intelligent systems and advocates for so-called 'world models' that learn from visual and spatial data to understand physical reality and causal relationships. He also indicates that current LLMs cannot retain and apply learnt information contextually across interactions (but see Yax et al. [8]).

Expanding this critique, several researchers argue that intelligence cannot be fully understood or replicated without considering the dynamic interplay between agents and their environments. For instance, Pedreschi et al. [68] suggest that intelligent behaviour would emerge from coevolutionary feedback loops between humans and ML systems. Rather than evolving separately, AI systems and their users mutually influence each other over time, shaping collective behaviour and potentially giving rise to unintended systemic outcomes. This view supports the notion that language, though central to human cognitive development, is embedded in broader architectures shaped by embodied interaction and social cooperation. Therefore, efforts to build synthetic intelligence must go beyond language and include sensorimotor feedback [356, 362] and motivational systems [66, 363] that mirror the integrative complexity of the human mind.

These limitations are consistent with perspectives from embodied cognition, which emphasise that cognitive processes emerge from the continuous coupling between perception, action and environment rather than from disembodied symbolic or statistical manipulation. Recent work on morphological computation provides a mechanistic account of this claim by showing how physical morphology itself performs computational work that supports and constrains sensing, control and learning across multiple levels of organisation [364]. In biological systems, and increasingly in robotics, body morphology functions as a bridge between the environment and internal control systems, effectively offloading part of the computational burden to physical dynamics, material properties and situated interaction with the world [365]. While the broader theoretical claims of embodied cognition remain debated [366], the practical implications for AI development are clear: Progress towards more human-like artificial cognition requires not only learning from multiple modalities beyond text and static vision but also embodied interaction. In embodied systems, learning is shaped by sensorimotor feedback, physical constraints and engagement with dynamic environments rather than by purely statistical correlations.

Another avenue for exploring the emergence of intelligence is artificial life (ALife) simulations. Rather than relying on top-down engineering, ALife seeks to recreate the ecological and interactive conditions under which intelligent behaviour might arise organically. Ramírez-Vizcaya and Froese [367] argue that these simulations provide a framework for accelerated, open-ended evolution within computational environments, where artificial agents gradually develop through self-organisation processes and interaction. This approach shifts the focus from preprogrammed architectures to emergent properties that may arise in multiagent ecosystems. Digital evolution platforms enable researchers to conduct experiments that would be intractable in biological systems by compressing thousands of generations into hours rather than years [368]. These systems have shown how complex cognitive capabilities can emerge through evolutionary processes without being explicitly programmed. Such experiments offer insights into the conditions necessary for the open-ended evolution of intelligence [369]. However, it is important to avoid anthropomorphising AI; superficial similarities in behaviour or learning patterns do not guarantee equivalence in understanding, intentionality or conscious experience [10, 240]. Such perspectives invite a broader reconsideration of how synthetic intelligence may unfold when designed to evolve rather than merely to predict.

9 | Conclusion

AI systems can improve their cognitive capabilities, efficiency and resilience through human-inspired cognitive mechanisms; however, Moravec's paradox is a reminder of disparities between human and AI capabilities. This principle, first articulated by robotics researcher Moravec [370], posits that high-level reasoning requires relatively little computation from AI systems, whereas sensorimotor and perceptual tasks demand enormous computational resources. Tasks that humans find effortless (e.g.,

recognising objects from different angles or manipulating objects with accurate movements) prove extraordinarily challenging for AI systems. Conversely, complex mathematical calculations or playing chess at the grandmaster level, activities that humans consider intellectually very demanding, are relatively straightforward for AI to master. Buckner [128] argues that this inverse relationship between human cognitive effort and AI computational requirements suggests that evolution has heavily tuned our neural circuitry for sensorimotor tasks, making them appear simple despite their objective complexity. Orhan and Lake [371] note that despite advances in DL, sensors and robotics, embodied AI remains a major challenge that encompasses sensorimotor skills and contextual understanding (see also Zador et al. [362]).

More broadly, the cognitive parallels documented in this review illustrate a recurring historical pattern in debates about AI. AI systems have progressively achieved cognitive functions once assumed to be uniquely human, such as expert reasoning (e.g., grandmaster-level chess), strategic planning under uncertainty (e.g., Go, poker) and creative production (e.g., visual art, music). Yet these achievements are frequently dismissed as 'surface intelligence' or mere pattern matching, which prompts new benchmarks of intelligence for AI to achieve and forces us to rethink the nature of human intelligence.

The rapid advancement of AI technologies can be deeply uncomfortable for many people, including researchers, even triggering feelings of being threatened or overwhelmed as machines appear to surpass human cognitive capacities [58]. Public perceptions of AI are often shaped less by an accurate understanding of technical capabilities than by social constructions infused with hopes, fears and cultural narratives [30, 64]. The present analysis suggests that much of the perceived threat and cycles of redefining benchmarks stem from long-standing overestimations of human cognitive capabilities. Moving beyond oversimplistic comparisons reveals a continuum of cognitive capabilities across human and artificial systems [100] (see also Steyvers et al. [351]). Both human and artificial systems are adaptive, probabilistic and shaped by learnt patterns, albeit within distinct constraints: biological in humans and computational in AI [293].

Even though understanding human cognition can inform AI development, we should be cautious about reverse-engineering human cognitive processes, given their limitations, biases and vulnerabilities [372]. Instead, the goal should be to understand both the strengths and weaknesses of human cognition to develop AI systems that can complement human capabilities. We should acknowledge that human cognition is not fully transparent and is subject to biases such as the illusion of explanatory depth [338], just as AI explanations often oversimplify the complexity of underlying models [331]. Our comparative narrative review has challenged traditional dichotomies between human and artificial cognition. The future of intelligence may not be defined by competition between human and artificial systems, but rather through a collaborative approach, requiring interdisciplinary dialogue between cognitive science, philosophy, system design and other disciplines of social and natural sciences.

Acknowledgements

The authors thank Delphine de Hemptinne for her helpful comments on earlier versions of the manuscript. During manuscript preparation, the authors used AI-assisted tools (OpenAI ChatGPT-4 and Anthropic Claude v3) to explore alternative formulations of interpretations and suggest phrasing options. All outputs from these tools were critically reviewed, verified and validated by all coauthors, who take full responsibility for the accuracy, integrity and final form of the manuscript. The authors also acknowledge the editorial support provided through Wiley Editing Services, which included two rounds of proofreading, narrative editing and reference cross-checking.

Funding

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant Program (Grant No. RGPIN-2022-04852).

Conflicts of Interest

The authors declare no conflicts of interest.

Endnotes

[1]The classification of fictional AI characters such as HAL 9000 (2001: A Space Odyssey), Data (Star Trek: The Next Generation) and C-3PO (Star Wars) as ANI, AGI or ASI is subject to ongoing debate among fans and bloggers (see, e.g., Myrow [373]). Interpretations often vary depending on factors such as autonomous strategic reasoning, emotional processing, learning flexibility, goal (mis)alignment and the scope of domain expertise (see Hermann [29], for a discussion of AI in metaphorical fiction).

[2]AI silver bullet is a common expression often used by researchers, media writers, bloggers and practitioners to refer to the unrealistic expectation that AI can solve all problems perfectly and effortlessly.

References

1. E. Hazan, A. Madgavkar, M. Chui, et al., *A New Future of Work: The Race to Deploy AI and Raise Skills in Europe and Beyond* (McKinsey Global Institute, 2024).

2. Valoir, *Assessing the Value of AI and Automation [Report]* (Valoir Inc, 2023), https://www.valoir.com.

3. IBM Corporation, *IBM Global AI Adoption Index 2022* (IBM, 2022), https://www.snowdropsolution.com/pdf/IBM%20Global%20AI%20Adoption%20Index%202022.pdf.

4. Gartner, "Gartner Poll Finds More Than Half of Organizations Have Increased Generative AI Investment in the Last 10 Months," (2023), https://www.gartner.com/en/newsroom/press-releases/2023-10-03-gartner-poll-finds-55-percent-of-organizations-are-in-piloting-or-production-mode-with-generative-ai.

5. M. Passalacqua, R. Pellerin, F. Magnani, et al., "Human-Centred AI in Industry 5.0: A Systematic Review," *International Journal of Production Research* 63, no. 7 (2025): 2638–2669, https://doi.org/10.1080/00207543.2024.2406021.

6. N. Kundu, F. Mustafa, K. Hemachandran, and C. Chola, "Artificial Intelligence in Retail Marketing," in *Artificial Intelligence for Business: An Implementation Guide Containing Practical and Industry-Specific Case Studies*, 1st ed., ed. K. Dans and R. Rodriguez (London: Routledge, 2023), 86–107, https://doi.org/10.4324/9781003358411.

7. P. C. Lazarus, P. E. Adeniyi, A. J. Ajayi, and D. M. Ajeyemi, "Harnessing Deep Learning for Advanced Visual Systems: Revolutionizing Computer Vision and Autonomous Navigation," *IRE Journal* 8, no. 2 (2024): 352–359.

8. N. Yax, H. Anlló, and S. Palminteri, "Studying and Improving Reasoning in Humans and Machines," *Communications Psychology* 2, no. 1 (2024): 51, https://doi.org/10.1038/s44271-024-00091-8.

9. F. Sense, R. Wood, M. G. Collins, et al., "Cognition-Enhanced Machine Learning for Better Predictions With Limited Data," *Topics in Cognitive Science* 14, no. 4 (2022): 739–755, https://doi.org/10.1111/tops.12574.

10. E. M. Bender, T. Gebru, A. McMillan-Major, and M. Mitchell, "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 2021)* (Toronto, 2021), 610–623.

11. N. Chater and M. Oaksford, *The Probabilistic Mind: Prospects for Bayesian Cognitive Science* (Oxford University Press, 2008).

12. D. C. Knill and A. Pouget, "The Bayesian Brain: The Role of Uncertainty in Neural Coding and Computation," *Trends in Neurosciences* 27, no. 12 (2004): 712–719, https://doi.org/10.1016/j.tins.2004.10.007.

13. D. C. Dennett and D. C. Dennett, *Consciousness Explained* (Penguin UK, 1993).

14. D. C. Dennett, *From Bacteria to Bach and Back: The Evolution of Minds* (W. W. Norton & Company, 2017).

15. S. Athey, "Beyond Prediction: Using Big Data for Policy Problems," *Science* 355, no. 6324 (2017): 483–485, https://doi.org/10.1126/science.aal4321.

16. M. N. Naveed Uddin, "Cognitive Science and Artificial Intelligence: Simulating the Human Mind and Its Complexity," *Cognitive Computation and Systems* 1, no. 4 (2019): 113–116, https://doi.org/10.1049/ccs.2019.0022.

17. A. Bundy, N. Chater, and S. Muggleton, "Introduction to Cognitive Artificial Intelligence," *Philosophical transactions. Series A, Mathematical, Physical, and Engineering Sciences* 381, no. 2251 (2023): 20220051, https://doi.org/10.1098/rsta.2022.0051.

18. M. K. Sarker, L. Zhou, A. Eberhart, and P. Hitzler, "Neuro-Symbolic Artificial Intelligence: Current Trends," *AI Communications* 34, no. 3 (2021): 197–209, https://doi.org/10.48550/arxiv.2105.05330.

19. N. Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014).

20. J. Abramson, J. Adler, J. Dunger, et al., "Accurate Structure Prediction of Biomolecular Interactions With Alphafold 3," *Nature* 630, no. 8016 (2024): 493–500, https://doi.org/10.1038/s41586-024-07487-w.

21. M. Kosinski, L. Wang, and R. Chen, "Evaluating GPT-4's Linguistic Capabilities," *Computational Linguistics* 50, no. 1 (2024): 23–42.

22. L. Fraade-Blanar, F. Favarò, J. Engstrom, et al., "Being Good (At Driving): Characterizing Behavioral Expectations on Automated and Human Driven Vehicles," (2025), https://arxiv.org/abs/2502.08121.

23. M. Adedjouma, B. Botella, J. Ibanez-Guzman, K. Mantissa, C.-M. Proum, and A. Smaoui, "Defining Operational Design Domain for Autonomous Systems: A Domain-Agnostic and Risk-Based Approach," in *Proceedings of the 19th Annual System of Systems Engineering Conference (SoSE)* (Tacoma, WA, 2024), 166–171, https://doi.org/10.1109/SOSE62659.2024.10620936.

24. R. Raman, R. Kowalski, K. Achuthan, A. Iyer, and P. Nedungadi, "Navigating Artificial General Intelligence Development: Societal, Technological, Ethical, and Brain-Inspired Pathways," *Scientific Reports* 15, no. 1 (2025): 8443, https://doi.org/10.1038/s41598-025-92190-7.

25. Y. LeCun, Y. Bengio, and G. Hinton, "Path Towards Autonomous Machine Intelligence," *Proceedings of the National Academy of Sciences of the United States of America* 120, no. 5 (2023): e2210630119.

26. K. Mahowald, A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko, "Dissociating Language and Thought in Large Language Models," *Trends in Cognitive Sciences* 28, no. 6 (2024): 517–540, https://doi.org/10.1016/j.tics.2024.01.011.

27. I. A. Blank, "What Are Large Language Models Supposed to Model?" *Trends in Cognitive Sciences* 27, no. 11 (2023): 987–989, https://doi.org/10.1016/j.tics.2023.08.006.

28. L. Dung, "Current Cases of AI Misalignment and Their Implications for Future Risks," *Synthese* 202, no. 5 (2023): 138, https://doi.org/10.1007/s11229-023-04367-0.

29. I. Hermann, "Artificial Intelligence in Fiction: Between Narratives and Metaphors," *AI & Society* 38, no. 1 (2023): 319–329, https://doi.org/10.1007/s00146-021-01299-6.

30. L. Sartori and G. Bocca, "Minding the Gap(S): Public Perceptions of AI and Socio-Technical Imaginaries," *AI & Society* 38, no. 2 (2023): 443–458, https://doi.org/10.1007/s00146-022-01422-1.

31. J. Taylor, E. Yudkowsky, P. LaVictoire, and A. Critch, "Alignment for Advanced Machine Learning Systems," in *Ethics of Artificial Intelligence*, ed. S. M. Liao (Oxford: Oxford University Book Company, 2020), 342–382, https://doi.org/10.1093/oso/9780190905033.003.0013.

32. E. Yudkowsky, "Pausing AI Developments Isn't Enough. We Need to Shut It all Down. Time Magazine," (2023), https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/.

33. C. Holl, "The Content Intelligence: An Argument Against the Lethality of Artificial Intelligence," *Discover Artificial Intelligence* 4, no. 1 (2024): 13, https://doi.org/10.1007/s44163-024-00112-9.

34. X. Dastile, T. Celik, and M. Potsane, "Statistical and Machine Learning Models in Credit Scoring: A Systematic Literature Survey," *Applied Soft Computing* 91 (2020): 106263, https://doi.org/10.1016/j.asoc.2020.106263.

35. Z. S. Wong, J. Zhou, and Q. Zhang, "Artificial Intelligence for Infectious Disease Surveillance and Outbreak Detection: A Systematic Review," *Journal of Biomedical Informatics* 139 (2023): 104298, https://doi.org/10.1016/j.jbi.2023.104298.

36. S. Lyu, X. Lang, H. Zhao, H. Zhang, P. Ding, and D. Wang, "RL2AC: Reinforcement Learning-Based Rapid Online Adaptive Control for Legged Robot Robust Locomotion," in *Proceedings of the 2024 Robotics: Science and Systems Conference* (Delft, 2024).

37. A. Narayanan and S. Kapoor, *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference* (Princeton University Press, 2024).

38. G. Orrù, M. Monaro, C. Conversano, A. Gemignani, and G. Sartori, "Machine Learning in Psychometrics and Psychological Research," *Frontiers in Psychology* 10 (2020): 2970, https://doi.org/10.3389/fpsyg.2019.02970.

39. D. A. Fife and J. D'Onofrio, "Common, Uncommon, and Novel Applications of Random Forest in Psychological Research," *Behavior Research Methods* 55, no. 5 (2023): 2447–2466, https://doi.org/10.3758/s13428-022-01901-9.

40. I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative Adversarial Networks," *Communications of the ACM* 63, no. 11 (2014): 139–144, https://doi.org/10.1145/3422622.

41. A. Ramesh, M. Pavlov, G. Goh, et al., "Zero-Shot Text-to-Image Generation," (2021), https://arxiv.org/abs/2102.12092.

42. J. Achiam, S. Adler, S. Agarwal, et al., "GPT-4 Technical Report," (2023), https://arxiv.org/abs/2303.08774.

43. R. Srinivasan, "Misinformation and Disinformation in Generative AI—A Survey," in *Proceedings of the 2025 Pacific-Asia Conference on Knowledge Discovery and Data Mining* (Sydney: Springer, 2025), 290–307.

44. M. R. Endsley, "From Here to Autonomy: Lessons Learned From Human-Automation Research," *Human Factors* 59, no. 1 (2017): 5–27, https://doi.org/10.1177/0018720816681350.

45. R. J. De Boer and S. W. Dekker, "Models of Automation Surprise: Results of a Field Survey in Aviation," *Safety Now* 3, no. 3 (2017): 20, https://doi.org/10.3390/safety3030020.

46. B. Kirwan, "Human Factors Requirements for Human AI Teaming in Aviation," *Future Transportation* 5, no. 2 (2025): 42, https://doi.org/10.3390/futuretransp5020042.

47. G. Dulac-Arnold, N. Levine, D. J. Mankowitz, et al., "Challenges of Real-World Reinforcement Learning: Definitions, Benchmarks and Analysis," *Machine Learning* 110, no. 9 (2021): 2419–2468, https://doi.org/10.1007/s10994-021-05961-4.

48. F. Miao and M. Cukurova, *AI Competency Framework for Teachers: Competencies for AI Literacy and Ethical Considerations in Education* (United Nations Educational, Scientific and Cultural Organization, 2024), https://www.unesco.org/en/articles/ai-competency-framework-teachers.

49. J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative Filtering Recommender Systems," in *The Adaptive Web*, ed. P. Brusilovsky, A. Kobsa, and W. Nejdl, 4321 (Berlin: Springer, 2007), 291–324.

50. Y. Li, K. Liu, R. Satapathy, S. Wang, and E. Cambria, "Recent Developments in Recommender Systems: A Survey," *IEEE Computational Intelligence Magazine* 19, no. 2 (2024): 78–95, https://doi.org/10.1109/MCI.2024.3363984.

51. M. Jesse and D. Jannach, "Digital Nudging With Recommender Systems: Survey and Future Directions," *Computers in Human Behavior Reports* 3 (2021): 100052, https://doi.org/10.1016/j.chbr.2020.100052.

52. K. Mei, X. Zhu, W. Xu, et al., "AIOS: LLM Agent Operating System," (2024), https://arxiv.org/abs/2403.16971.

53. C. Burns, P. Izmailov, J. H. Kirchner, et al., "Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision," (2023), https://arxiv.org/abs/2312.09390.

54. Q. Wu, G. Bansal, J. Zhang, et al., "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation," (2023), https://arxiv.org/abs/2308.08155.

55. G. Li, H. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "Camel: Communicative Agents for mind Exploration of Large Language Model Society," *Advances in Neural Information Processing Systems NeurIPS conference* 36 (New Orleans, USA, 2023): 51991–52008.

56. B. Marr, "The Third Wave of AI Is Here: Why Agentic AI Will Transform the Way We Work," (2024), https://www.forbes.com/sites/bernardmarr/2024/02/26/the-third-wave-of-ai-is-here-why-agentic-ai-will-transform-the-way-we-work/.

57. G. E. Gignac and E. T. Szodorai, "Defining Intelligence: Bridging the Gap Between Human and Artificial Perspectives," *Intelligence* 104 (2024): 101832, https://doi.org/10.1016/j.intell.2024.101832.

58. J. E. Korteling, G. C. van de Boer-Visschedijk, R. A. M. Blankendaal, R. C. Boonekamp, and A. R. Eikelboom, "Human-Versus Artificial Intelligence," *Frontiers in Artificial Intelligence* 4 (2021): https://doi.org/10.3389/frai.2021.622364.

59. J. M. Bishop, "Artificial Intelligence is Stupid and Causal Reasoning Will Not Fix It," *Frontiers in Psychology* 11 (2020): 513474, https://doi.org/10.3389/fpsyg.2020.513474.

60. D. McDermott, "Level-Headed," *Artificial Intelligence* 171, no. 18 (2007): 1183–1186, https://doi.org/10.1016/j.artint.2007.10.013.

61. C. Buckner, "Black Boxes or Unflattering Mirrors? Comparative Bias in the Science of Machine Behaviour," *The British Journal for the Philosophy of Science* 74, no. 3 (2023): 681–712, https://doi.org/10.1086/714960.

62. J. Suomala and J. Kauttonen, "Human's Intuitive Mental Models as a Source of Realistic Artificial Intelligence and Engineering," *Frontiers in Psychology* 13 (2022): 873289, https://doi.org/10.3389/fpsyg.2022.873289.

63. J. Zerilli, A. Knott, J. Maclaurin, and C. Gavaghan, "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?" *Philosophy & Technology* 32, no. 4 (2019): 661–683, https://doi.org/10.1007/s13347-018-0330-6.

64. D. K. Kanbach, L. Heiduk, G. Blueher, M. Schreiter, and A. Lahmann, "The GenAI is Out of the Bottle: Generative Artificial Intelligence From a Business Model Innovation Perspective," *Review of Managerial Science* 18, no. 4 (2024): 1189–1220, https://doi.org/10.1007/s11846-023-00696-z.

65. J. Zhao, M. Wu, L. Zhou, X. Wang, and J. Jia, "Cognitive Psychology-Based Artificial Intelligence Review," *Frontiers in Neuroscience* 16 (2022): 1024316, https://doi.org/10.3389/fnins.2022.1024316.

66. S. Pinker, *Language, Cognition, and Human Nature* (Oxford University Press, 2013).

67. Y. LeCun, "AI is Not intelligent—Yet," (2024), https://www.newsweek.com/ai-impact-interview-yann-lecun-llm-limitations-analysis-2054255.

68. D. Pedreschi, L. Pappalardo, E. Ferragina, et al., "Human-AI Coevolution," *Artificial Intelligence* 339 (2025): 104244, https://doi.org/10.1016/j.artint.2024.104244.

69. R. Ferrari, "Writing Narrative Style Literature Reviews," *Medical Writing* 24, no. 4 (2015): 230–235, https://doi.org/10.1179/2047480615Z.000000000329.

70. T. Greenhalgh, S. Thorne, and K. Malterud, "Time to Challenge the Spurious Hierarchy of Systematic Over Narrative Reviews?" *European Journal of Clinical Investigation* 48, no. 6 (2018): e12931, https://doi.org/10.1111/eci.12931.

71. U. Neisser, *Cognitive Psychology* (Appleton-Century-Crofts, 1967).

72. U. Neisser, *Cognition and Reality: Principles and Implications of Cognitive Psychology* (W. H. Freeman, 1976).

73. A. Newell and H. A. Simon, *Human Problem Solving*, 104 (Prentice-Hall, 1972).

74. D. Marr, *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information* (W. H. Freeman, 1982).

75. H. Simon, "Cognitive Science: The Newest Science of the Artificial," *Cognitive Science* 4, no. 1 (1980): 33–46, https://doi.org/10.1016/s0364-0213(81)80003-1.

76. H. A. Simon, "Information-Processing Models of Cognition," *Journal of the American Society for Information Science* 32, no. 5 (1981): 364–377, https://doi.org/10.1002/asi.4630320517.

77. G. A. Miller, "The Cognitive Revolution: A Historical Perspective," *Trends in Cognitive Sciences* 7, no. 3 (2003): 141–144, https://doi.org/10.1016/S1364-6613(03)00029-9.

78. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning Internal Representations by Error Propagation* (MIT Press, 1986).

79. B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building Machines That Learn and Think Like People," *Behavioral and Brain Sciences* 40 (2017): e253, https://doi.org/10.1017/S0140525X16001837.

80. J. E. T. Taylor and G. W. Taylor, "Artificial Cognition: How Experimental Psychology Can Help Generate Explainable Artificial Intelligence," *Psychonomic Bulletin & Review* 28, no. 2 (2021): 454–475, https://doi.org/10.3758/s13423-020-01825-5.

81. J. Moor, "The Dartmouth College Artificial Intelligence Conference: The next Fifty Years," *AI Magazine* 27, no. 4 (2006): 87.

82. A. Newell and H. A. Simon, *The Simulation of Human Thought* (Rand Corporation, 1959).

83. A. Newell, "The Knowledge Level," *Artificial Intelligence* 18, no. 1 (1982): 87–127, https://doi.org/10.1016/0004-3702(82)90012-1.

84. M. McCloskey, "Networks and Theories: the Place of Connectionism in Cognitive Science," in *Cognitive Modeling*, ed. T. Polk and C. Seifert (Cambridge, MA: The MIT Press, 2002), 1131–1148, https://doi.org/10.7551/mitpress/1888.003.0042.

85. M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry* (Cambridge Press, 1969).

86. J. L. McClelland, D. E. Rumelhart, and G. E. Hinton, "The Appeal of Parallel Distributed Processing," in *Cognitive Psychology*, ed. D. Balota and E. J. Marsh (Cambridge, MA: MIT Press, 1986), 3–44.

87. G. E. Hinton, S. Osindero, and Y.-W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Computation* 18, no. 7 (2006): 1527–1554, https://doi.org/10.1162/neco.2006.18.7.1527.

88. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification With Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 25 (Cambridge, MA: MIT Press, 2012).

89. A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is All You Need," (2023), https://arxiv.org/abs/1706.03762.

90. L. Serafini, I. Donadello, and A. D. A. Garcez, "Learning and Reasoning in Logic Tensor Networks: Theory and Application to Semantic Image Interpretation," in *Proceedings of the 2017 Symposium on Applied Computing* (Marrakech, 2017), 125–130.

91. D. Marr, "Artificial intelligence—A Personal View," *Artificial Intelligence* 9, no. 1 (1977): 37–48, https://doi.org/10.1016/0004-3702(77)90013-3.

92. P. Quinlan, "Marr's Vision 30 Years On: From a Personal Point of View," *Perception* 41, no. 9 (2012): 1009–1012, https://doi.org/10.1068/p4109ed.

93. J. McCarthy, "Recursive Functions of Symbolic Expressions and Their Computation by Machine, Part I," *Communications of the ACM* 3, no. 4 (1960): 184–195, https://doi.org/10.1145/367177.367199.

94. D. N. Cassenti, V. D. Veksler, and F. E. Ritter, "Editor's Review and Introduction: Cognition-Inspired Artificial Intelligence," *Topics in Cognitive Science* 14, no. 4 (2022): 652–664, https://doi.org/10.1111/tops.12622.

95. T. Bayne, A. K. Seth, M. Massimini, et al., "Tests for Consciousness in Humans and Beyond," *Trends in Cognitive Sciences* 28, no. 5 (2024): 454–466, https://doi.org/10.1016/j.tics.2024.01.010.

96. E. M. Bender, "Resisting Dehumanization in the Age of AI," *Current Directions in Psychological Science* 33, no. 2 (2024): 114–120, https://doi.org/10.1177/09637214231217286.

97. E. Favier-Baron, "Que Fait L'Intelligence Artificielle à L'Intelligence?" *Appareil* 26 (2023): https://doi.org/10.4000/appareil.6943.

98. R. J. Sternberg, "Do Not Worry That Generative AI May Compromise Human Creativity or Intelligence in the Future: It Already Has," *Journal of Intelligence* 12, no. 7 (2024): 69, https://doi.org/10.3390/jintelligence12070069.

99. K. F. Hubert, K. N. Awa, and D. L. Zabelina, "The Current State of Artificial Intelligence Generative Language Models is More Creative Than Humans on Divergent Thinking Tasks," *Scientific Reports* 14, no. 1 (2024): 3440, https://doi.org/10.1038/s41598-024-53303-w.

100. G. Siemens, F. Marmolejo-Ramos, F. Gabriel, et al., "Human and Artificial Cognition," *Computers and Education: Artificial Intelligence* 3 (2022): 100107, https://doi.org/10.1016/j.caeai.2022.100107.

101. A. Gilchrist, *Seeing Black and White* (Oxford University Press, 2006).

102. J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman, "How to Grow a Mind: Statistics, Structure, and Abstraction," *Science* 331, no. 6022 (2011): 1279–1285, https://doi.org/10.1126/science.1192788.

103. J. M. Findlay and I. D. Gilchrist, *Active Vision: The Psychology of Looking and Seeing (No. 37)* (Oxford University Press, 2003).

104. N. Cowan, "The Magical Mystery Four: How is Working Memory Capacity Limited, and Why?" *Current Directions in Psychological Science* 19, no. 1 (2010): 51–57, https://doi.org/10.1177/0963721409359277.

105. N. Lavie, "Attention, Distraction, and Cognitive Control Under Load," *Current Directions in Psychological Science* 19, no. 3 (2010): 143–148, https://doi.org/10.1177/0963721410370295.

106. C. Wickens, "Attention: Theory, Principles, Models and Applications," *International Journal of Human-Computer Interaction* 37, no. 5 (2021): 403–417, https://doi.org/10.1080/10447318.2021.1874741.

107. B. Anderson, "There Is No Such Thing as Attention," *Frontiers in Psychology* 2 (2011): 246, https://doi.org/10.3389/fpsyg.2011.00246.

108. B. Hommel, C. S. Chapman, P. Cisek, H. F. Neyedli, J.-H. Song, and T. N. Welsh, "No One Knows What Attention is," *Attention, Perception, & Psychophysics* 81, no. 7 (2019): 2288–2303, https://doi.org/10.3758/s13414-019-01846-w.

109. D. L. Schacter and D. R. Addis, "The Cognitive Neuroscience of Constructive Memory: Remembering the Past and Imagining the Future," *Philosophical Transactions of the Royal Society B* 362, no. 1481 (2007): 773–786, https://doi.org/10.1098/rstb.2007.2087.

110. E. Tulving, "Episodic and Semantic Memory," in *Organization of Memory*, ed. E. Tulving and W. Donaldson (Cambridge, MA: Academic Press, 1972), 381–403.

111. P. N. Johnson-Laird and E. Shafir, "The Interaction Between Reasoning and Decision Making: An Introduction," *Cognition* 49, no. 1-2 (1993): 1–9, https://doi.org/10.1016/0010-0277(93)90033-R.

112. L. E. Guerrero, L. F. Castillo, J. Arango-López, and F. Moreira, "A Systematic Review of Integrated Information Theory: A Perspective From Artificial Intelligence and the Cognitive Sciences," *Neural Computing & Applications* 37, no. 11 (2025): 7575–7607, https://doi.org/10.1007/s00521-023-08328-z.

113. C. Gonzalez, "Building Human-Like Artificial Agents: A General Cognitive Algorithm for Emulating Human Decision-Making in Dynamic Environments," *Perspectives on Psychological Science* 19, no. 5 (2024): 860–873, https://doi.org/10.1177/17456916231196766.

114. T. V. P. Bliss and T. Lømo, "Long-Lasting Potentiation of Synaptic Transmission in the Dentate Area of the Anaesthetized Rabbit Following Stimulation of the Perforant Path," *The Journal of Physiology* 232, no. 2 (1973): 331–356, https://doi.org/10.1113/jphysiol.1973.sp010273.

115. D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory* (John Wiley & Sons, 1949).

116. N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, and R. J. Dolan, "Model-Based Influences on Humans' Choices and Striatal Prediction Errors," *Neuron* 69, no. 6 (2011): 1204–1215, https://doi.org/10.1016/j.neuron.2011.02.027.

117. S. Carey, *The Origin of Concepts* (Oxford University Press, 2009).

118. B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-Level Concept Learning Through Probabilistic Program Induction," *Science* 350, no. 6266 (2015): 1332–1338, https://doi.org/10.1126/science.aab3050.

119. M. T. H. Chi, N. de Leeuw, M.-H. Chiu, and C. LaVancher, "Eliciting Self-Explanations Improves Understanding," *Cognitive Science* 18, no. 3 (1994): 439–477, https://doi.org/10.1207/s15516709cog1803_3.

120. T. Lombrozo, "Learning by Thinking in Natural and Artificial Minds," *Trends in Cognitive Sciences* 28, no. 11 (2024): 1011–1022, https://doi.org/10.1016/j.tics.2024.07.007.

121. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proceedings of the 2020 International Conference on Machine Learning* (Vienna, 2020), 1597–1607.

122. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Minneapolis, MN, 2019), 4171–4186.

123. V. Mnih, K. Kavukcuoglu, D. Silver, et al., "Human-Level Control Through Deep Reinforcement Learning," *Nature* 518, no. 7540 (2015): 529–533, https://doi.org/10.1038/nature14236.

124. R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. (MIT Press, 2018).

125. C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," in *Proceedings of the 2017 International Conference on Machine Learning* (Sydney, 2017), 1126–1135.

126. A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-Learning With Memory-Augmented Neural Networks," in *Proceedings of the 2016 International Conference on Machine Learning* (New York, NY, 2016), 1842–1850.

127. J. Von Oswald, E. Niklasson, E. Randazzo, et al., "Transformers Learn In-Context by Gradient Descent," in *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, HI, 2023).

128. C. Buckner, *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us About the Future of Artificial Intelligence* (Oxford University Press, 2024).

129. D. L. Hintzman, J. J. Summers, and R. A. Block, "What Causes the Spacing Effect? Some Effects of Repetition, Duration, and Spacing on Memory for Pictures," *Memory & Cognition* 3, no. 3 (1975): 287–294, https://doi.org/10.3758/BF03212913.

130. R. Bhui and R. Dubey, "Why Context Should Matter," *Decision* 11, no. 4 (2024): 557–567, https://doi.org/10.1037/dec0000234.

131. J. X. Wang, "Meta-Learning in Natural and Artificial Intelligence," *Current Opinion in Behavioral Sciences* 38 (2021): 90–95, https://doi.org/10.1016/j.cobeha.2021.01.002.

132. Y. Liu, J. Ma, Y. Xie, et al., "Contrastive Predictive Coding With Transformer for Video Representation Learning," *Neurocomputing* 482 (2022): 154–162, https://doi.org/10.1016/j.neucom.2021.11.031.

133. T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large Language Models Are Zero-Shot Reasoners," in *Advances in Neural Information Processing Systems*, 35 (Cambridge, MA: MIT Press, 2022), 22199–22213.

134. J. Wei, X. Wang, D. Schuurmans, et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," (2022), https://arxiv.org/abs/2201.11903.

135. R. M. Sapolsky, *Determined: Life Without Free Will* (Random House, 2023).

136. C. Kidd, S. T. Piantadosi, and R. N. Aslin, "The Goldilocks Effect in Infant Auditory Attention," *Child Development* 85, no. 5 (2014): 1795–-1804, https://doi.org/10.1111/cdev.12263.

137. A. N. Sanborn and N. Chater, "Bayesian Brains Without Probabilities," *Trends in Cognitive Sciences* 20, no. 12 (2016): 883–893, https://doi.org/10.1016/j.tics.2016.10.003.

138. P. Johansson, L. Hall, and S. Sikström, "From Change Blindness to Choice Blindness," *Psychologia* 51, no. 2 (2008): 142–155, https://doi.org/10.2117/psysoc.2008.142.

139. W. S. Geisler and R. L. Diehl, "A Bayesian Approach to the Evolution of Perceptual and Cognitive Systems," *Cognitive Science* 27, no. 3 (2003): 379–402, https://doi.org/10.1207/s15516709cog2703_3.

140. K. J. Friston, "A Theory of Cortical Responses," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 360, no. 1456 (2005): 815–836, https://doi.org/10.1098/rstb.2005.1622.

141. A. Clark, "Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science," *Behavioral and Brain Sciences* 36, no. 3 (2013): 181–204, https://doi.org/10.1017/S0140525X12000477.

142. Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature* 521, no. 7553 (2015): 436–444, https://doi.org/10.1038/nature14539.

143. D. Silver, A. Huang, C. J. Maddison, et al., "Mastering the Game of Go With Deep Neural Networks and Tree Search," *Nature* 529, no. 7587 (2016): 484–489, https://doi.org/10.1038/nature16961.

144. J. Li, Z. Yu, Z. Du, L. Zhu, and H. T. Shen, "A Comprehensive Survey on Source-Free Domain Adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, no. 8 (2024): 5743–5762, https://doi.org/10.1109/TPAMI.2024.3370978.

145. A. Radford, J. W. Kim, C. Hallacy, et al., "Learning Transferable Visual Models From Natural Language Supervision," (2021), https://arxiv.org/abs/2103.00020.

146. E. H. Adelson and J. R. Bergen, "The Plenoptic Function and the Elements of Early Vision," in *Computational Models of Visual Processing*, ed. M. S. Landy and J. A. Movshon (Cambridge, MA: MIT Press, 1991), 3–20, https://doi.org/10.7551/mitpress/2002.003.0004.

147. I. Biederman, "Recognition-by-Components: A Theory of Human Image Understanding," *Psychological Review* 94, no. 2 (1987): 115–147, https://doi.org/10.1037/0033-295X.94.2.115.

148. D. H. Hubel and T. N. Wiesel, "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex," *The Journal of Physiology* 160, no. 1 (1962): 106–154, https://doi.org/10.1113/jphysiol.1962.sp006837.

149. T. Vladusich and M. D. McDonnell, "A Unified Account of Perceptual Layering and Surface Appearance in Terms of Gamut Relativity," *PLoS One* 9, no. 11 (2014): e113159, https://doi.org/10.1371/journal.pone.0113159.

150. A. Gilchrist and A. Soranzo, "What is the Relationship Between Lightness and Perceived Illumination?" *Journal of Experimental Psychology: Human Perception and Performance* 45, no. 11 (2019): 1470–-1483, https://doi.org/10.1037/xhp0000675.

151. S. Ullman, "Visual Routines," *Cognition* 18, no. 1–3 (1984): 97–159, https://doi.org/10.1016/0010-0277(84)90023-4.

152. F. A. Wichmann and R. Geirhos, "Are Deep Neural Networks Adequate Behavioral Models of Human Visual Perception?" *Annual Review of Vision Science* 9, no. 1 (2023): 501–524, https://doi.org/10.1146/annurev-vision-120522-031739.

153. A. Nayebi, J. Sagastuy-Brena, D. M. Bear, et al., "Recurrent Connections in the Primate Ventral Visual Stream Mediate a Trade-Off Between Task Performance and Network Size During Core Object Recognition," *Neural Computation* 34, no. 8 (2022): 1652–1675, https://doi.org/10.1162/neco_a_01506.

154. P. Agrawal, J. Carreira, and J. Malik, "In What Do Deep Networks See? Adapting Retinal Sampling for Robust Visual Processing," in *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Vancouver, 2023), 2172–2181.

155. H. Lin, K. Zhang, and Y. Tian, "Perceptual Diffusion: Integrating Human-Inspired Attention With Vision Transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, no. 1 (2024): 112–125.

156. A. de Santana Correia and E. L. Colombini, "Attention, Please! A Survey of Neural Attention Models in Deep Learning," *Artificial Intelligence Review* 55, no. 8 (2022): 6037–6124, https://doi.org/10.1007/s10462-022-10148-x.

157. R. F. Murray, "Lightness Perception in Complex Scenes," *Annual Review of Vision Science* 7, no. 1 (2021): 417–436, https://doi.org/10.1146/annurev-vision-093019-115159.

158. J. Feather, G. Leclerc, A. Mądry, and J. H. McDermott, "Model Metamers Reveal Divergent Invariances Between Biological and Artificial Neural Networks," *Nature Neuroscience* 26, no. 11 (2023): 2017–2034, https://doi.org/10.1038/s41593-023-01442-0.

159. J. Freeman and E. P. Simoncelli, "Metamers of the Ventral Stream," *Nature Neuroscience* 14, no. 9 (2011): 1195–1201, https://doi.org/10.1038/nn.2889.

160. D. D. Salvucci and N. A. Taatgen, "Threaded Cognition: An Integrated Theory of Concurrent Multitasking," *Psychological Review* 115, no. 1 (2008): 101–130, https://doi.org/10.1037/0033-295X.115.1.101.

161. R. Fischer and F. Plessow, "Efficient Multitasking: Parallel Versus Serial Processing of Multiple Tasks," *Frontiers in Psychology* 6 (2015): 1366, https://doi.org/10.3389/fpsyg.2015.01366.

162. A. Kiesel, M. Steinhauser, M. Wendt, et al., "Control and Interference in Task Switching—A Review," *Psychological Bulletin* 136, no. 5 (2010): 849–874, https://doi.org/10.1037/a0019842.

163. R. Thakur, N. Goël, N. Bhavesh, and V. Vijaykumar, "Enhancing Deep Learning Performance Through Parallel Processing: A Comprehensive Research Study," in *Proceedings of the 6th International Conference on Recent Trends in Advance Computing (ICRTAC)* (Chennai, 2023), 836–841, https://doi.org/10.1109/ICRTAC59277.2023.10480769.

164. H. M. Hodgetts, S. Packwood, F. Vachon, and S. Tremblay, "A Microworld Simulation of Dynamic Cognition as a Test of Executive Function," *Journal of Clinical and Experimental Neuropsychology* 45, no. 2 (2023): 165–181, https://doi.org/10.1080/13803395.2023.2214297.

165. R. F. Adler and R. Benbunan-Fich, "Juggling on a High Wire: Multitasking Effects on Performance," *International Journal of Human-Computer Studies* 70, no. 2 (2012): 156–168, https://doi.org/10.1016/j.ijhcs.2011.10.003.

166. M. Krichen and M. S. Abdalzaher, "Performance Enhancement of Artificial Intelligence: A Survey," *Journal of Network and Computer Applications* 232 (2024): 104034, https://doi.org/10.1016/j.jnca.2024.104034.

167. G. Menghani, "Efficient Deep Learning: A Survey on Making Deep Learning Models Smaller, Faster, and Better," *ACM Computing Surveys* 55, no. 12 (2023): 1–37, https://doi.org/10.1145/3578938.

168. I. Kotseruba and J. K. Tsotsos, "40 Years of Cognitive Architectures: Core Cognitive Abilities and Practical Applications," *Artificial Intelligence Review* 53, no. 1 (2020): 17–94, https://doi.org/10.1007/s10462-018-9646-y.

169. L. Ding, J. Terwilliger, A. Parab, et al., "CLERA: A Unified Model for Joint Cognitive Load and Eye Region Analysis in the Wild," *ACM Transactions on Computer-Human Interaction* 30, no. 6 (2023): 1–23, https://doi.org/10.1145/3603622.

170. B. Reimer, B. Mehler, Y. Wang, and J. F. Coughlin, "A Field Study on the Impact of Variations in Short-Term Memory Demands on Drivers' Visual Attention and Driving Performance Across Three Age Groups," *Human Factors* 54, no. 3 (2012): 454–468, https://doi.org/10.1177/0018720812437274.

171. E. Merlo, E. Lamon, F. Fusaro, et al., "An Ergonomic Role Allocation Framework for Dynamic Human-Robot Collaborative Tasks," *Journal of Manufacturing Systems* 67 (2023): 111–121, https://doi.org/10.1016/j.jmsy.2022.12.011.

172. B. A. Richards, T. P. Lillicrap, P. Beaudoin, et al., "A Deep Learning Framework for Neuroscience," *Nature Neuroscience* 22, no. 11 (2019): 1761–1770, https://doi.org/10.1038/s41593-019-0520-2.

173. J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, "Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights From the Successes and Failures of Connectionist Models of Learning and Memory," *Psychological Review* 102, no. 3 (1995): 419–457, https://doi.org/10.1037/0033-295X.102.3.419.

174. E. R. Kandel, Y. Dudai, and M. R. Mayford, "The Molecular and Systems Biology of Memory," *Cell* 157, no. 1 (2014): 163–186, https://doi.org/10.1016/j.cell.2014.03.001.

175. J. T. Wixted, "The Role of Retroactive Interference and Consolidation in Everyday Forgetting," in *Current Issues in Memory*, ed. J. Rummel (London: Routledge, 2021), 117–143, https://doi.org/10.4324/9781003106715-8.

176. L. Nadel and M. Moscovitch, "Memory Consolidation, Retrograde Amnesia and the Hippocampal Complex," *Current Opinion in Neurobiology* 7, no. 2 (1997): 217–227, https://doi.org/10.1016/s0959-4388(97)80010-4.

177. R. M. French, "Catastrophic Forgetting in Connectionist Networks," *Trends in Cognitive Sciences* 3, no. 4 (1999): 128–135, https://doi.org/10.1016/s1364-6613(99)01294-2.

178. T. McGrath, A. Kapishnikov, N. Tomašev, et al., "Acquisition of Chess Knowledge in Alphazero," *Proceedings of the National Academy of Sciences of the United States of America* 119, no. 47 (2022): e2206625119, https://doi.org/10.1073/pnas.2206625119.

179. T. Tadros, G. P. Krishnan, R. Ramyaa, and M. Bazhenov, "Sleep-Like Unsupervised Replay Reduces Catastrophic Forgetting in Artificial Neural Networks," *Nature Communications* 13, no. 1 (2022): 7742, https://doi.org/10.1038/s41467-022-34938-7.

180. J. Kirkpatrick, R. Pascanu, N. Rabinowitz, et al., "Overcoming Catastrophic Forgetting in Neural Networks," *Proceedings of the National Academy of Sciences of the United States of America* 114, no. 13 (2017): 3521–3526, https://doi.org/10.1073/pnas.1611835114.

181. D. Lopez-Paz and M. Ranzato, "Gradient Episodic Memory for Continual Learning," *Advances in Neural Information Processing Systems*, 30 (Cambridge, MA: MIT Press, 2017).

182. E. F. Loftus, "Planting Misinformation in the Human Mind: A 30-Year Investigation of the Malleability of Memory," *Learning & Memory* 12, no. 4 (2005): 361–366, https://doi.org/10.1101/lm.94705.

183. E. F. Loftus and J. C. Palmer, "Reconstruction of Automobile Destruction: An Example of the Interaction Between Language and Memory," *Journal of Verbal Learning and Verbal Behavior* 13, no. 5 (1974): 585–589, https://doi.org/10.1016/S0022-5371(74)80011-3.

184. M. C. Anderson and S. Hanslmayr, "Neural Mechanisms of Motivated Forgetting," *Trends in Cognitive Sciences* 18, no. 6 (2014): 279–292, https://doi.org/10.1016/j.tics.2014.03.002.

185. M. A. Conway, "Memory and the Self," *Journal of Memory and Language* 53, no. 4 (2005): 594–628, https://doi.org/10.1016/j.jml.2005.08.005.

186. E. A. Kensinger, "Remembering the Details: Effects of Emotion," *Emotion Review* 1, no. 2 (2009): 99–113, https://doi.org/10.1177/1754073908100432.

187. J. De Jong, S. Wilhelm, and E. G. Akyürek, "Adaptive Forgetting Speed in Working Memory," *Psychonomic Bulletin & Review* 31, no. 6 (2024): 2704–2713, https://doi.org/10.3758/s13423-024-02507-2.

188. J. R. Anderson and R. Milson, "Human Memory: An Adaptive Perspective," *Psychological Review* 96, no. 4 (1989): 703–719, https://doi.org/10.1037/0033-295X.96.4.703.

189. R. A. Bjork and E. L. Bjork, "A New Theory of Disuse and an Old Theory of Stimulus Fluctuation. From Learning Processes to Cognitive Processes," *Essays in Honor of William K. Estes*, 2 (London: Routledge, 1992), 35–67.

190. E. T. Cowan, A. C. Schapiro, J. E. Dunsmoor, and V. P. Murty, "Memory Consolidation as an Adaptive Process," *Psychonomic Bulletin & Review* 28, no. 6 (2021): 1796–1810, https://doi.org/10.3758/s13423-021-01978-x.

191. A. P. Yonelinas and M. Ritchey, "The Slow Forgetting of Emotional Episodic Memories: An Emotional Binding Account," *Trends in Cognitive Sciences* 19, no. 5 (2015): 259–267, https://doi.org/10.1016/j.tics.2015.02.009.

192. D. Shohamy and R. A. Adcock, "Dopamine and Adaptive Memory," *Trends in Cognitive Sciences* 14, no. 10 (2010): 464–472, https://doi.org/10.1016/j.tics.2010.08.002.

193. K. Murayama and S. Kitagami, "Consolidation Power of Extrinsic Rewards: Reward Cues Enhance Long-Term Memory for Irrelevant Past Events," *Journal of Experimental Psychology: General* 143, no. 1 (2014): 15–20, https://doi.org/10.1037/a0031992.

194. M. Mitchell and D. C. Krakauer, "The Debate Over Understanding in AI's Large Language Models," *Proceedings of the National Academy of*

Sciences 120, no. 13 (2023): e2215907120, https://doi.org/10.1073/pnas.2215907120.

195. B. A. Richards and P. W. Frankland, "The Persistence and Transience of Memory," *Neuron* 94, no. 6 (2017): 1071–1084, https://doi.org/10.1016/j.neuron.2017.04.037.

196. Y. Hu, S. Liu, Y. Yue, et al., "Memory in the Age of AI Agents: A Survey," (2026), https://arxiv.org/abs/2512.13564.

197. M. Riemer, I. Cases, R. Ajemian, et al., "Learning to Learn Without Forgetting by Maximizing Transfer and Minimizing Interference," (2018), https://arxiv.org/abs/1810.11910.

198. L. Lavoie-Hudon, C. Bureau, D. Lafond, and S. Tremblay, "Less Can Be More: Effects of a Forgetting Function on an AI-Based Policy Capturing Tool Performance," *Proceedings of the Human Factors and Ergonomics Society-Annual Meeting* 69, no. 1 (2025): 1601–1607.

199. W. Edwards, "Dynamic Decision Theory and Probabilistic Information Processings," *Human Factors: The Journal of the Human Factors and Ergonomics Society* 4, no. 2 (1962): 59–74, https://doi.org/10.1177/001872086200400201.

200. D. Kahneman, *Thinking, Fast and Slow* (Farrar, Straus and Giroux, 2011).

201. A. Tversky and D. Kahneman, "Judgment Under Uncertainty: Heuristics and Biases," *Science* 185, no. 4157 (1974): 1124–1131, https://doi.org/10.1126/science.185.4157.1124.

202. C. Gonzalez, P. Fakhari, and J. Busemeyer, "Dynamic Decision Making: Learning Processes and New Research Directions," *Human Factors* 59, no. 5 (2017): 713–721, https://doi.org/10.1177/0018720817710347.

203. D. Dörner and J. Funke, "Complex Problem Solving: What It Is and What It Is Not," *Frontiers in Psychology* 8 (2017): 1153, https://doi.org/10.3389/fpsyg.2017.01153.

204. A. Nachbagauer, "Managing Complexity in Projects: Extending the Cynefin Framework," *Project Leadership and Society* 2 (2021): 100017, https://doi.org/10.1016/j.plas.2021.100017.

205. D. J. Snowden and M. E. Boone, "A Leader's Framework for Decision Making," *Harvard Business Review* 85, no. 11 (2007): 68–76.

206. J. Funke, "Complex Problem Solving," *Encyclopedia of the Sciences of Learning* (Berlin: Springer, 2012), 682–685.

207. J. S. B. T. Evans and K. E. Stanovich, "Dual-Process Theories of Higher Cognition: Advancing the Debate," *Perspectives on Psychological Science* 8, no. 3 (2013): 223–241, https://doi.org/10.1177/1745691612460685.

208. P. Bory, "Deep New: The Shifting Narratives of Artificial Intelligence From Deep Blue to Alphago," *Convergence* 25, no. 4 (2019): 627–642, https://doi.org/10.1177/1354856519829679.

209. A. Ananthaswamy, "Deepmind AI Topples Experts at Complex Game Stratego," *Nature* (2022): https://doi.org/10.1038/d41586-022-04246-7.

210. N. Brown and T. Sandholm, "Superhuman AI for Multiplayer Poker," *Science* 365, no. 6456 (2019): 885–890, https://doi.org/10.1126/science.aay2400.

211. O. Vinyals, I. Babuschkin, W. M. Czarnecki, et al., "Grandmaster Level in Starcraft II Using Multi-Agent Reinforcement Learning," *Nature* 575, no. 7782 (2019): 350–354, https://doi.org/10.1038/s41586-019-1724-z.

212. T. Hagendorff, S. Fabi, and M. Kosinski, "Human-Like Intuitive Behavior and Reasoning Biases Emerged in Large Language Models But Disappeared in Chatgpt," *Nature Computational Science* 3, no. 10 (2023): 833–838, https://doi.org/10.1038/s43588-023-00527-x.

213. M. Binz and E. Schulz, "Using Cognitive Psychology to Understand GPT-3," *Proceedings of the National Academy of Sciences of the United States of America* 120, no. 6 (2023): e2218523120, https://doi.org/10.1073/pnas.2218523120.

214. D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, "GPT-4 Passes the Bar Exam," *Philosophical Transactions of the Royal Society a Mathematical Physical and Engineering Sciences* 382 (2023): 2270, https://doi.org/10.1098/rsta.2023.0254.

215. G. Marcus and E. Davis, "Insights for AI From the Human Mind," *Communications of the ACM* 64, no. 1 (2021): 38–41, https://doi.org/10.1145/3392663.

216. H. A. Simon, "Bounded Rationality in Social Science: Today and Tomorrow," *Mind & Society* 1, no. 1 (2000): 25–39, https://doi.org/10.1007/BF02512227.

217. M. A. Goodrich, W. C. Stirling, and E. R. Boer, "Satisficing Revisited," *Minds and Machines* 10, no. 1 (2000): 79–109, https://doi.org/10.1023/A:1008325423033.

218. S. Hélie and Z. Pizlo, "When Is Psychology Research Useful in Artificial Intelligence? A Case for Reducing Computational Complexity in Problem Solving," *Topics in Cognitive Science* 14, no. 4 (2021): 687–701, https://doi.org/10.1111/tops.12572.

219. G. Gigerenzer, "The Bias in Behavioral Economics," *Review of Behavioral Economics* 5, no. 3-4 (2018): 303–336, https://doi.org/10.1561/105.00000092.

220. F. Lieder and T. L. Griffiths, "Resource-Rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources," *Behavioral and Brain Sciences* 43 (2020): e1, https://doi.org/10.1017/S0140525X1900061X.

221. K. V. Katsikopoulos, Ö. Şimşek, M. Buckmann, and G. Gigerenzer, *Classification in the Wild: The Science and Art of Transparent Decision Making* (MIT Press, 2021).

222. G. Gigerenzer and H. Brighton, "Homo Heuristicus: Why Biased Minds Make Better Inferences," *Topics in Cognitive Science* 1, no. 1 (2009): 107–143, https://doi.org/10.1111/j.1756-8765.2008.01006.x.

223. R. Bhui, L. Lai, and S. J. Gershman, "Resource-Rational Decision Making," *Current Opinion in Behavioral Sciences* 41 (2021): 15–21, https://doi.org/10.1016/j.cobeha.2021.02.015.

224. D. Dörner and C. D. Güss, "Human Error in Complex Problem Solving and Dynamic Decision Making: A Taxonomy of 24 Errors and a Theory," *Computers in Human Behavior Reports* 7 (2022): 100222, https://doi.org/10.1016/j.chbr.2022.100222.

225. A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics Derived Automatically From Language Corpora Contain Human-Like Biases," *Science* 356, no. 6334 (2017): 183–186, https://doi.org/10.1126/science.aal4230.

226. J. K. Nguyen, "Human Bias in AI Models? Anchoring Effects and Mitigation Strategies in Large Language Models," *Journal of Behavioral and Experimental Finance* 43 (2024): 100971, https://doi.org/10.1016/j.jbef.2024.100971.

227. M. H. Christiansen and N. Chater, "The Now-or-Never Bottleneck: A Fundamental Constraint on Language," *Behavioral and Brain Sciences* 39 (2016): e62, https://doi.org/10.1017/S0140525X1500031X.

228. B. Lake and M. Baroni, "Generalization Without Systematicity: on the Compositional Skills of sequence-to-sequence Recurrent Networks," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, ed. J. Dy and A. Krause (Stockholm, 2018), 4487–4499.

229. G. Thierry, "We Need to Stop Pretending AI Is Intelligent – Here's How," (2025), https://theconversation.com/we-need-to-stop-pretending-ai-is-intelligent-heres-how-254090.

230. W. R. Uttal, *Dualism* (Routledge, 2004).

231. F. Verbruggen, I. P. L. McLaren, and C. D. Chambers, "Banishing the Control Homunculi in Studies of Action Control and Behavior

Change," *Perspectives on Psychological Science* 9, no. 5 (2014): 497–524, https://doi.org/10.1177/1745691614526414.

232. T. E. Hazy, M. J. Frank, and R. C. O'Reilly, "Banishing the Homunculus: Making Working Memory Work," *Neuroscience* 139, no. 1 (2006): 105–118, https://doi.org/10.1016/j.neuroscience.2005.04.067.

233. J. Margolis, "The Trouble With Homunculus Theories," *Philosophy of Science* 47, no. 2 (1980): 244–259, https://doi.org/10.1086/288931.

234. G. Thierry, C. D. Martin, P. E. Downing, and A. J. Pegna, "Controlling for Interstimulus Perceptual Variance Abolishes N170 Face Selectivity," *Nature Neuroscience* 10, no. 4 (2007): 505–511, https://doi.org/10.1038/nn1864.

235. B. J. Baars, "The Philosophical Mind-Body Debate: Why It's Still With Us," (2024), https://bernardbaars.substack.com/p/the-philosophical-mind-body-debate.

236. S. Dehaene, H. Lau, and S. Kouider, "What Is Consciousness, and Could Machines Have It?" in *Robotics, AI, and Humanity*, ed. J. von Braun, S. Archer, G. M. Reichberg, M. Sánchez Sorondo, and M. Sánchez Sorondo (Berlin: Springer, 2021), 43–56, https://doi.org/10.1007/978-3-030-54173-6_4.

237. A. K. Seth, "Conscious Artificial Intelligence and Biological Naturalism," *Behavioral and Brain Sciences* (2025): 1–42, https://doi.org/10.1017/S0140525X25000032.

238. T. Felin and M. Holweg, "Theory Is All You Need: AI, Human Cognition, and Causal Reasoning," *Strategy Science* 9, no. 4 (2024): 346–371, https://doi.org/10.1287/stsc.2024.0189.

239. T. Felin, J. Koenderink, and J. I. Krueger, "Rationality, Perception, and the All-Seeing Eye," *Psychonomic Bulletin & Review* 24, no. 4 (2017): 1040–1059, https://doi.org/10.3758/s13423-016-1198-z.

240. J. R. Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3, no. 3 (1980): 417–424, https://doi.org/10.1017/s0140525x00005756.

241. N. Chater, *The Mind is Flat: The Remarkable Shallowness of the Improvising Brain* (Yale University Press, 2018).

242. H. A. Simon, "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics* 69, no. 1 (1955): 99, https://doi.org/10.2307/1884852.

243. F. C. Keil, "Explanation and Understanding," *Annual Review of Psychology* 57, no. 1 (2006): 227–254, https://doi.org/10.1146/annurev.psych.57.102904.190100.

244. T. Lombrozo and N. Vasilyeva, "Causal Explanation," in *The Oxford Handbook of Causal Reasoning*, ed. M. R. Waldmann (Oxford: Oxford University Press, 2017), 415–432.

245. R. Moreno-Bote, D. C. Knill, and A. Pouget, "Bayesian Sampling in Visual Perception," *Proceedings of the National Academy of Sciences* 108, no. 30 (2011): 12491–12496, https://doi.org/10.1073/pnas.1101430108.

246. S. Khemlani and P. N. Johnson-Laird, "Theories of the Syllogism: A Meta-Analysis," *Psychological Bulletin* 138, no. 3 (2012): 427–457, https://doi.org/10.1037/a0026841.

247. N. Riesterer, D. Brand, H. Dames, and M. Ragni, "Modeling Human Syllogistic Reasoning: The Role of No Valid Conclusion," *Topics in Cognitive Science* 12, no. 1 (2020): 446–459, https://doi.org/10.1111/tops.12487.

248. J. S. B. Evans, L. J. Ball, and V. A. Thompson, "Belief Bias in Deductive Reasoning," in *Cognitive Illusions*, ed. R. F. Pohl (London: Routledge, 2022), 154–172.

249. P. N. Johnson-Laird, "Mental Models and Human Reasoning," *Proceedings of the National Academy of Sciences of the United States of America* 107, no. 43 (2010): 18243–18250, https://doi.org/10.1073/pnas.1012933107.

250. M. Oaksford and N. Chater, "Précis of Bayesian Rationality: The Probabilistic Approach to Human Reasoning," *Behavioral and Brain Sciences* 32, no. 1 (2009): 69–84, https://doi.org/10.1017/S0140525X09000284.

251. J. H. Holland, "Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology," *Control, and Artificial Intelligence* (Cambridge, MA: MIT Press, 1992), https://doi.org/10.7551/mitpress/1090.001.0001.

252. T. L. Griffiths, N. Chater, and J. B. Tenenbaum, *Bayesian Models of Cognition: Reverse Engineering the Mind* (MIT Press, 2024).

253. K. Stenning and M. van Lambalgen, *Human Reasoning and Cognitive Science* (MIT Press, 2008).

254. M. Ebrahimi, A. Eberhart, F. Bianchi, and P. Hitzler, "Towards Bridging the Neuro-Symbolic Gap: Deep Deductive Reasoners," *Applied Intelligence* 51, no. 9 (2021): 6326–6348, https://doi.org/10.1007/s10489-020-02165-6.

255. M. H. Tessler, J. B. Tenenbaum, and N. D. Goodman, "Logic, Probability, and Pragmatics in Syllogistic Reasoning," *Topics in Cognitive Science* 14, no. 3 (2022): 574–601, https://doi.org/10.1111/tops.12593.

256. S. J. Han, K. J. Ransom, A. Perfors, and C. Kemp, "Inductive Reasoning in Humans and Large Language Models," *Cognitive Systems Research* 83 (2024): 101155, https://doi.org/10.1016/j.cogsys.2023.101155.

257. R. A. Fabio, D. Verzì, and A. Gangemi, "A Contribution to the Default-Interventionist and Parallel Accounts in Deductive Reasoning: The Effect of Decisional Styles on Logic and Belief," *The Journal of General Psychology* 151, no. 2 (2024): 209–222, https://doi.org/10.1080/00221309.2023.2241952.

258. K. Cheng, J. Yang, H. Jiang, et al., "Inductive or Deductive? Rethinking the Fundamental Reasoning Abilities of LLMs," (2024), https://arxiv.org/abs/2408.00114.

259. B. Jiang, Y. Xie, Z. Hao, et al., "A Peek Into Token Bias: Large Language Models Are Not Yet Genuine Reasoners," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Miami, FL: Association for Computational Linguistics, 2024), 4722–-4756, https://doi.org/10.18653/v1/2024.emnlp-main.272.

260. K. Hamilton, A. Nayak, B. Božić, and L. Longo, "Is Neuro-Symbolic AI Meeting Its Promises in Natural Language Processing? A Structured Review," *Semantic Web* 15, no. 4 (2024): 1265–1306, https://doi.org/10.3233/SW-223228.

261. A. D. Garcez and L. C. Lamb, "Neurosymbolic AI: The 3rd Wave," (2020), https://arxiv.org/abs/2012.05876.

262. Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving Factuality and Reasoning in Language Models Through Multiagent Debate," in *Proceedings of the 41st International Conference on Machine Learning* (Vienna, 2023).

263. T. Liang, Z. He, W. Jiao, et al., "Encouraging Divergent Thinking in Large Language Models Through Multi-Agent Debate," (2023), https://arxiv.org/abs/2305.19118.

264. S. Blackmore, "Why Good Explanations Matter," *New Scientist* 244, no. 3259 (2019): 26–27, https://doi.org/10.1016/S0262-4079(19)32313-9.

265. R. Rousseau, S. Tremblay, S. Banbury, R. Breton, and A. Guitouni, "The Role of Metacognition in the Relationship Between Objective and Subjective Measures of Situation Awareness," *Theoretical Issues in Ergonomics Science* 11, no. 1-2 (2010): 119–130, https://doi.org/10.1080/14639220903010076.

266. S. Katyal and S. M. Fleming, "The Future of Metacognition Research: Balancing Construct Breadth With Measurement Rigor," *Cortex* 171 (2024): 223–234, https://doi.org/10.1016/j.cortex.2023.11.002.

267. D. C. Dennett, *Intuition Pumps and Other Tools for Thinking* (W. W. Norton & Company, 2013).

268. J. Haidt, "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," *Psychological Review* 108, no. 4 (2001): 814–834, https://doi.org/10.1037/0033-295x.108.4.814.

269. G. Pennycook, R. M. Ross, D. J. Koehler, and J. A. Fugelsang, "Dunning -Kruger Effects in Reasoning: Theoretical Implications of the Failure to Recognize Incompetence," *Psychonomic Bulletin & Review* 24, no. 6 (2017): 1774–1784, https://doi.org/10.3758/s13423-017-1242-7.

270. J. Krueger, R. A. Mueller, and Unaware Unskilled, "or Both? The Better-Than-Average Heuristic and Statistical Regression Predict Errors in Estimates of Own Performance," *Journal of Personality and Social Psychology* 82, no. 2 (2002): 180–188, https://doi.org/10.1037/0022-3514.82.2.180.

271. S. Lichtenstein and P. Slovic, *The Construction of Preference* (Cambridge University Press, 2006).

272. T. D. Wilson and D. T. Gilbert, "Affective Forecasting: Knowing What to Want," *Current Directions in Psychological Science* 14, no. 3 (2005): 131–134, https://doi.org/10.1111/j.0963-7214.2005.00355.x.

273. R. E. Nisbett and T. D. Wilson, "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* 84, no. 3 (1977): 231–259, https://doi.org/10.1037/0033-295X.84.3.231.

274. K. A. Ericsson, "Protocol Analysis," in *A Companion to Cognitive Science*, ed. W. Bechtel and G. Graham (Hoboken, NJ: Blackwell Publishing Ltd, 2017), 425–432.

275. S. M. Fleming and R. J. Dolan, "The Neural Basis of Metacognitive Ability," *Philosophical Transactions of the Royal Society B* 367, no. 1594 (2012): 1338–1349, https://doi.org/10.1098/rstb.2011.0417.

276. J. Metcalfe and A. P. Shimamura, *Metacognition: Knowing About Knowing* (MIT Press, 1994).

277. F. J. Binder, J. Chua, T. Korbak, et al., "Looking Inward: Language Models Can Learn About Themselves by Introspection," (2024), https://arxiv.org/abs/2410.13787.

278. V. Kartha, "Self-Reflecting AI Agents Using Langchain," (2023), https://vijaykumarkartha.medium.com/self-reflecting-ai-agents-using-langchain-d3a93684da92.

279. L. A. Renier, M. Schmid Mast, and A. Bekbergenova, "To Err is Human, Not Algorithmic – Robust Reactions to Erring Algorithms," *Computers in Human Behavior* 124 (2021): 106879, https://doi.org/10.1016/j.chb.2021.106879.

280. R. L. Gregory, "Knowledge in Perception and Illusion," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 352, no. 1358 (1997): 1121–1127, https://doi.org/10.1098/rstb.1997.0095.

281. J. Hohwy, "The Predictive Processing Hypothesis," in *The Oxford Handbook of 4E Cognition*, ed. A. Newen, L. De Bruin, and S. Gallagher (Oxford: Oxford Library of Psychology, 2018), 129–145.

282. D. M. Eagleman, "Visual Illusions and Neurobiology," *Nature Reviews Neuroscience* 2, no. 12 (2001): 920–926, https://doi.org/10.1038/35104092.

283. D. Deutsch, "The Paradox of Pitch Circularity," *Acoustics Today* 7, no. 3 (2010): 8–15, https://doi.org/10.1121/1.3488670.

284. R. M. Warren, "Perceptual Restoration of Missing Speech Sounds," *Science* 167, no. 3917 (1970): 392–393, https://doi.org/10.1126/science.167.3917.392.

285. J. Metcalfe, "Learning From Errors," *Annual Review of Psychology* 68, no. 1 (2017): 465–489, https://doi.org/10.1146/annurev-psych-010416-044022.

286. X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial Examples: Attacks and Defenses for Deep Learning," *IEEE Transactions on Neural Networks and Learning Systems* 30, no. 9 (2019): 2805–2824, https://doi.org/10.1109/TNNLS.2018.2886017.

287. G. Pennycook and D. G. Rand, "The Psychology of Fake News," *Trends in Cognitive Sciences* 25, no. 5 (2021): 388–402, https://doi.org/10.1016/j.tics.2021.02.007.

288. S. Hooker, "Moving Beyond Algorithmic Bias is a Data Problem," *Patterns* 2, no. 4 (2021): 100241, https://doi.org/10.1016/j.patter.2021.100241.

289. D. Hendrycks, S. Basart, N. Mu, et al., "The Many Faces of Robustness: a Critical Analysis of out-of-distribution Generalization," in *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision* (Montreal, 2021), 8340–8349.

290. J. Rasmussen, "Human Errors. A Taxonomy for Describing Human Malfunction in Industrial Installations," *Journal of Occupational Accidents* 4, no. 2–4 (1982): 311–333, https://doi.org/10.1016/0376-6349(82)90041-4.

291. B. Schneier and N. Sanders, "AI Mistakes Are Very Different From Human Mistakes," (2025), https://spectrum.ieee.org/ai-mistakes-schneier.

292. M. Liu, J. Wei, Y. Liu, and J. Davis, "Human and AI Perceptual Differences in Image Classification Errors," (2025), https://arxiv.org/abs/2304.08733.

293. C. Novelli, F. Casolari, A. Rotolo, M. Taddeo, and L. Floridi, "Taking AI Risks Seriously: A New Assessment Model for the AI Act," *AI & Society* 39, no. 5 (2024): 2493–2497, https://doi.org/10.1007/s00146-023-01723-z.

294. D. Ariely, *Predictably Irrational* (Harper Collins, 2008).

295. P. F. Nichelli and J. Grafman, "The Place of Free Will: The Freedom of the Prisoner," *Neurological Sciences* 45, no. 3 (2024): 861–871, https://doi.org/10.1007/s10072-023-07138-4.

296. B. Béchard, M. Ouimet, H. M. Hodgetts, F. Morneau-Guérin, and S. Tremblay, "Political Complexity and the Pervading Role of Ideology in Policy-Making," *Journal of Dynamic Decision Making* 9 (2024).

297. B. Libet, C. A. Gleason, E. W. Wright, and D. K. Pearl, "Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential). The Unconscious Initiation of a Freely Voluntary Act," *Brain* 106, no. 3 (1983): 623–642, https://doi.org/10.1093/brain/106.3.623.

298. M. Matsuhashi and M. Hallett, "The Timing of the Conscious Intention to Move," *European Journal of Neuroscience* 28, no. 11 (2008): 2344–2351, https://doi.org/10.1111/j.1460-9568.2008.06525.x.

299. D. M. Wegner and T. Wheatley, "Apparent Mental Causation. Sources of the Experience of Will," *American Psychologist* 54, no. 7 (1999): 480–492, https://doi.org/10.1037//0003-066x.54.7.480.

300. B. Kotchoubey, "Human Consciousness: Where is It From and What is It For," *Frontiers in Psychology* 9 (2018): 567, https://doi.org/10.3389/fpsyg.2018.00567.

301. M. Brass, A. Furstenberg, and A. R. Mele, "Why Neuroscience Does Not Disprove Free Will," *Neuroscience & Biobehavioral Reviews* 102 (2019): 251–263, https://doi.org/10.1016/j.neubiorev.2019.04.024.

302. L. Mudrik, I. G. Arie, Y. Amir, et al., "Free Will Without Consciousness?" *Trends in Cognitive Sciences* 26, no. 7 (2022): 555–566, https://doi.org/10.1016/j.tics.2022.03.005.

303. Q. Wang and A. K. Goel, "Mutual Theory of Mind for Human-AI Communication," (2024), https://arxiv.org/abs/2210.03842.

304. J. Williams, S. M. Fiore, and F. Jentsch, "Supporting Artificial Social Intelligence With Theory of Mind," *Frontiers in Artificial Intelligence* 5 (2022): 750763, https://doi.org/10.3389/frai.2022.750763.

305. S. M. Carlson, M. A. Koenig, and M. B. Harms, "Theory of Mind," *Cognitive Science* 4, no. 4 (2013): 391–402, https://doi.org/10.1002/wcs.1232.

306. M. Farisco, K. Evers, and J.-P. Changeux, "Is Artificial Consciousness Achievable? Lessons From the Human Brain," *Neural Networks* 180 (2024): 106714, https://doi.org/10.1016/j.neunet.2024.106714.

307. J. Aru, A. Labash, O. Corcoll, and R. Vicente, "Mind the Gap: Challenges of Deep Learning Approaches to Theory of Mind," *Artificial Intelligence Review* 56, no. 9 (2023): 9141–9156, https://doi.org/10.1007/s10462-023-10401-x.

308. U. Peters, "Explainable AI Lacks Regulative Reasons: Why AI and Human Decision-Making Are Not Equally Opaque," *AI and Ethics* 3, no. 3 (2023): 963–974, https://doi.org/10.1007/s43681-022-00217-w.

309. S. Russell, *Human Compatible: AI and the Problem of Control* (Penguin UK, 2019).

310. M. Botvinick, J. X. Wang, W. Dabney, K. J. Miller, and Z. Kurth-Nelson, "Deep Reinforcement Learning and Its Neuroscientific Implications," *Neuron* 107, no. 4 (2020): 603–616, https://doi.org/10.1016/j.neuron.2020.06.014.

311. P. S. Churchland, *Brain-Wise: Studies in Neurophilosophy* (MIT Press, 2005).

312. A. Gopnik, "Artificial Minds and Developmental Models: What Can Children Teach Machines?" *Trends in Cognitive Sciences* 27, no. 1 (2023): 5–13.

313. R. Malach, "The Neuronal Basis of Human Creativity," *Frontiers in Human Neuroscience* 18 (2024): 1367922, https://doi.org/10.3389/fnhum.2024.1367922.

314. G. Fauconnier and M. Turner, *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities* (Basic Books, 2003).

315. N. Leach, "In the Mirror of AI: What Is Creativity?" *Architectural Intelligence* 1, no. 1 (2022): 15, https://doi.org/10.1007/s44223-022-00012-x.

316. D. Cropley, "Is Artificial Intelligence More Creative Than Humans? Chatgpt and the Divergent Association Task," *Learning Letters* 2 (2023): 13, https://doi.org/10.59453/ll.v2.13.

317. S. A. Mednick, "The Associative Basis of the Creative Process," *Psychological Review* 69, no. 3 (1962): 220–232, https://doi.org/10.1037/h0048850.

318. M. Benedek and A. C. Neubauer, "Revisiting Mednick's Model on Creativity-Related Differences in Associative Hierarchy: Evidence for a Common Path to Uncommon Thought," *Journal of Creative Behavior* 47, no. 4 (2013): 273–289, https://doi.org/10.1002/jocb.35.

319. C. Liu, K. Zhuang, D. C. Zeitlen, et al., "Neural, Genetic, and Cognitive Signatures of Creativity," *Communications Biology* 7, no. 1 (2024): 1324, https://doi.org/10.1038/s42003-024-07007-6.

320. R. Wingström, J. Hautala, and R. Lundman, "Redefining Creativity in the Era of AI? Perspectives of Computer Scientists and New Media Artists," *Creativity Research Journal* 36, no. 2 (2024): 177–193, https://doi.org/10.1080/10400419.2022.2107850.

321. J. Henrich, S. J. Heine, and A. Norenzayan, "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33, no. 2-3 (2010): 61–83, https://doi.org/10.1017/S0140525X0999152X.

322. G. Lupyan and A. Clark, "Words and the World: Predictive Coding and the Language-Perception-Cognition Interface," *Current Directions in Psychological Science* 24, no. 4 (2015): 279–284, https://doi.org/10.1177/0963721415570732.

323. T. Regier and Y. Xu, "The Sapir-Whorf Hypothesis and Inference Under Uncertainty," *Cognitive Science* 9, no. 3 (2017): e1464, https://doi.org/10.1002/wcs.1464.

324. L. Tredinnick and C. Laybats, "Black-Box Creativity and Generative Artificial Intelligence," *Business Information Review* 40, no. 3 (2023): 98–102, https://doi.org/10.1177/02663821231195131.

325. M. Ragot, N. Martin, and S. Cojean, "AI-Generated Vs. Human Artworks: A Perception Bias Towards Artificial Intelligence?" in *Proceedings of the Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, 2020), 1–10, https://doi.org/10.1145/3334480.3382892.

326. F. Magni, J. Park, and M. M. Chao, "Humans as Creativity Gatekeepers: Are We Biased Against AI Creativity?" *Journal of Business and Psychology* 39, no. 3 (2024): 643–656, https://doi.org/10.1007/s10869-023-09910-x.

327. S. Grassini and M. Koivisto, "Artificial Creativity? Evaluating AI Against Human Performance in Creative Interpretation of Visual Stimuli," *International Journal of Human-Computer Interaction* 41, no. 7 (2025): 4037–4048, https://doi.org/10.1080/10447318.2024.2345430.

328. E. Zhou and D. Lee, "Generative Artificial Intelligence, Human Creativity, and Art," *PNAS Nexus* 3, no. 3 (2024): 52, https://doi.org/10.1093/pnasnexus/pgae052.

329. P. R. Lewis and Ş. Sarkadi, "Reflective Artificial Intelligence," *Minds and Machines* 34, no. 2 (2024): 14, https://doi.org/10.1007/s11023-024-09664-2.

330. V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A Review of Trustworthy and Explainable Artificial Intelligence (XAI)," *IEEE Access* 11 (2023): 78994–79015, https://doi.org/10.1109/ACCESS.2023.3294569.

331. Z. C. Lipton, "The Mythos of Model Interpretability," *Communications of the ACM* 61, no. 10 (2018): 36–43, https://doi.org/10.1145/3233231.

332. E. Nasarian, R. Alizadehsani, U. R. Acharya, and K.-L. Tsui, "Designing Interpretable Ml System to Enhance Trust in Healthcare: A Systematic Review to Proposed Responsible Clinician-AI-Collaboration Framework," *Information Fusion* 108 (2024): 102412, https://doi.org/10.1016/j.inffus.2024.102412.

333. C. Rudin and J. Radin, "Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From an Explainable AI Competition," *Harvard Data Science Review* 1, no. 2 (2019): https://doi.org/10.1162/99608f92.5a8a3a3d.

334. D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G. Z. Yang, "XAI—Explainable Artificial Intelligence," *Science Robotics* 4, no. 37 (2019): eaay7120, https://doi.org/10.1126/scirobotics.aay7120.

335. G. Klein, "Naturalistic Decision Making," *Human Factors* 50, no. 3 (2008): 456–460, https://doi.org/10.1518/001872008X288385.

336. R. R. Hoffman, G. Klein, and S. T. Mueller, "Explaining Explanation for Explainable AI," *Proceedings of the Human Factors and Ergonomics Society-Annual Meeting* 62, no. 1 (2018): 197–201, https://doi.org/10.1177/1541931218621047.

337. A. Raz, B. Heinrichs, N. Avnoon, G. Eyal, and Y. Inbar, "Prediction and Explainability in AI: Striking a New Balance?" *Big Data and Society* 11, no. 1 (2024): 20539517241235871, https://doi.org/10.1177/20539517241235871.

338. L. Rozenblit and F. Keil, "The Misunderstood Limits of Folk Science: An Illusion of Explanatory Depth," *Cognitive Science* 26, no. 5 (2002): 521–562, https://doi.org/10.1207/s15516709cog2605_1.

339. Q. V. Liao and J. W. Vaughan, "AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap," *Harvard Data Science Review* 5, no. 5 (2024): https://doi.org/10.1162/99608f92.8036d03b.

340. S. Tremblay, J.-F. Gagnon, D. Lafond, H. M. Hodgetts, M. Doiron, and P. P. J. M. H. Jeuniaux, "A Cognitive Prosthesis for Complex Decision-Making," *Applied Ergonomics* 58 (2017): 349–360, https://doi.org/10.1016/j.apergo.2016.07.009.

341. L. Yan, S. Greiff, Z. Teuber, and D. Gašević, "Promises and Challenges of Generative Artificial Intelligence for Human Learning," *Nature Human Behaviour* 8, no. 10 (2024): 1839–1850, https://doi.org/10.1038/s41562-024-02004-5.

342. S. Baron, "Explainable AI and Causal Understanding: Counterfactual Approaches Considered," *Minds and Machines* 33, no. 2 (2023): 347–377, https://doi.org/10.1007/s11023-023-09637-x.

343. B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens, and J. Vanthienen, "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring," *Journal of the Operational Research Society* 54, no. 6 (2003): 627–635, https://doi.org/10.1057/palgrave.jors.2601545.

344. L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: an Overview of Interpretability of Machine Learning," in *Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (Turin: IEEE, 2018), 80–89, https://doi.org/10.1109/DSAA.2018.00018.

345. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, 30 (Cambridge, MA: MIT Press, 2017).

346. S. M. Lundberg, G. Erion, H. Chen, et al., "From Local Explanations to Global Understanding With Explainable AI for Trees," *Nature Machine Intelligence* 2, no. 1 (2020): 56–67, https://doi.org/10.1038/s42256-019-0138-9.

347. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA, 2016), 1135–1144, https://doi.org/10.1145/2939672.2939778.

348. B. Mittelstadt, C. Russell, and S. Wachter, "Explaining Explanations in AI," in *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, 2019), 279–288, https://doi.org/10.1145/3287560.3287574.

349. T. Miller, "Explanation in Artificial Intelligence: Insights From the Social Sciences," *Artificial Intelligence* 267 (2019): 1–38, https://doi.org/10.1016/j.artint.2018.07.007.

350. V. Hassija, V. Chamola, A. Mahapatra, et al., "Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence," *Cognitive Computation* 16, no. 1 (2024): 45–74, https://doi.org/10.1007/s12559-023-10179-8.

351. M. Steyvers, H. Tejeda, A. Kumar, et al., "What Large Language Models Know and What People Think They Know," *Nature Machine Intelligence* 7, no. 2 (2025): 221–231, https://doi.org/10.1038/s42256-024-00976-7.

352. D. Estes and K. Bartsch, "Theory of Mind: A Foundational Component of Human General Intelligence," *Behavioral and Brain Sciences* 40 (2017): e201, https://doi.org/10.1017/S0140525X16001618.

353. S. Pinker, "Colloquium Paper: The Cognitive Niche: Coevolution of Intelligence, Sociality, and Language," *Proceedings of the National Academy of Sciences of the United States of America* 107, no. Suppl. 2 (2010): 8993–8999, https://doi.org/10.1073/pnas.0914630107.

354. G. Lupyan, "The Centrality of Language in Human Cognition," *Language Learning* 66, no. 3 (2016): 516–553, https://doi.org/10.1111/lang.12155.

355. D. Gentner, "Language as Cognitive Tool Kit: How Language Supports Relational Thought," *American Psychologist* 71, no. 8 (2016): 650–657, https://doi.org/10.1037/amp0000082.

356. A. Clark, *Supersizing the Mind: Embodiment, Action, and Cognitive Extension* (Oxford University Press, 2010).

357. F. J. Varela, E. Rosch, and E. Thompson, *The Embodied Mind: Cognitive Science and Human Experience* (MIT Press, 1991).

358. P. Wolff and K. J. Holmes, "Linguistic Relativity," *Wiley Interdisciplinary Reviews. Cognitive Science* 2, no. 3 (2011): 253–265, https://doi.org/10.1002/wcs.104.

359. C. Li, Z. Gan, Z. Yang, et al., "Multimodal Foundation Models: From Specialists to General-Purpose Assistants," *Foundations and Trends® in Computer Graphics and Vision* 16, no. 1-2 (2024): 1–214, https://doi.org/10.1561/9781638283379.

360. H. Kaur, C. Kishor Kumar Reddy, D. Manoj Kumar Reddy, and M. M. Hanafiah, "Single Modality to Multi-Modality: The Evolutionary Trajectory of Artificial Intelligence in Integrating Diverse Data Streams for Enhanced Cognitive Capabilities," in *Multimodal Generative AI*, ed. A. Singh and K. K. Singh (Berlin: Springer, 2025), 297–322, https://doi.org/10.1007/978-981-96-2355-6_13.

361. B. Xu and M.-M. Poo, "Large Language Models and Brain-Inspired General Intelligence," *National Science Review* 10, no. 10 (2023): nwad267, https://doi.org/10.1093/nsr/nwad267.

362. A. Zador, S. Escola, B. Richards, et al., "Catalyzing Next-Generation Artificial Intelligence Through Neuroai," *Nature Communications* 14, no. 1 (2023): 1597, https://doi.org/10.1038/s41467-023-37180-x.

363. D. J. Jilk, C. Lebiere, R. C. O'Reilly, and J. R. Anderson, "SAL: An Explicitly Pluralistic Cognitive Architecture," *Journal of Experimental & Theoretical Artificial Intelligence* 20, no. 3 (2008): 197–218, https://doi.org/10.1080/09528130802319128.

364. G. Dodig-Crnkovic, "Cognition as Embodied Morphological Computation," in *Philosophy and Theory of Artificial Intelligence 2017*, ed. V. C. Müller, 44 (Berlin: Springer, 2018), 1–5, https://doi.org/10.1007/978-3-319-96448-5_2.

365. H. Hauser and J. Hughes, "Morphological Computation—Past, Present and Future," *Device* 2, no. 9 (2024): 100439, https://doi.org/10.1016/j.device.2024.100439.

366. S. D. Goldinger, M. H. Papesh, Barnhart, et al., "The Poverty of Embodied Cognition," *Psychonomic Bulletin & Review* 23, no. 4 (2016): 959–978, https://doi.org/10.3758/s13423-015-0860-1.

367. S. Ramírez-Vizcaya and T. Froese, "Agents of Habit: Refining the Artificial Life Route to Artificial Intelligence," *Artificial Life Conference Proceedings* 32 (2020): 78–86.

368. E. Dolson and C. Ofria, "Digital Evolution for Ecology Research: A Review," *Frontiers in Ecology and Evolution* 9 (2021): 750779, https://doi.org/10.3389/fevo.2021.750779.

369. L. Soros and K. Stanley, "Identifying Minimal Conditions for open-ended Evolution Through the Artificial Life World of Chromaria," in *Proceedings of the ALIFE 14: The 14th International Conference on the Synthesis and Simulation of Living Systems* (New York, NY: MIT Press, 2014), 793–800.

370. H. Moravec, *Mind Children: The Future of Robot and Human Intelligence* (Harvard University Press, 1988).

371. A. E. Orhan and B. M. Lake, "Learning High-Level Visual Representations From a Child's Perspective Without Strong Inductive Biases," *Nature Machine Intelligence* 6, no. 3 (2024): 271–283, https://doi.org/10.1038/s42256-024-00802-0.

372. O. Macmillan-Scott and M. Musolesi, "(Ir)rationality and Cognitive Biases in Large Language Models," *Royal Society Open Science* 11, no. 6 (2024): 240255, https://doi.org/10.1098/rsos.240255.

373. R. Myrow, "AI in the Movies. From HAL to "Her". KQED," (2023), https://www.kqed.org/arts/13932477/ai-movies-2001-matrix-colossus-star-trek-data-terminator-blade-runner.