Theory Is All You Need:
AI, Human Cognition, and Decision Making†

Teppo Felin
Utah State University
& Oxford University

Matthias Holweg
Oxford University

† Arguments related to this paper were presented at the *Strategy Science* "Theory-Based View" conference at Bocconi University, Harvard Business School, Aalto University, and the University of Illinois Urbana-Champaign. We are grateful for feedback from many participants and audience members that have helped us improve our arguments. We appreciate feedback from (or related conversations with) Gopesh Anand, Arnaldo Camuffo, Alfonso Gambardella, Pranav Gupta, Jared Hansen, Rosco Hunter, Sharad Jones, Stuart Kauffman, Jan Koenderink, Matt Kraatz, Karim Lakhani, Natalia Levina, Hila Lifshitz-Assaf, Jeff Loewenstein, Geoff Love, Jukka Luoma, Frank Martela, Mariano Mastrogiorgio, Joe Mahoney, Willie Ocasio, Samuli Reijula, Chris Rytting, Mari Sako, Jens Schmidt, Cirrus Shakeri, Olivier Sibony, Deepak Somaya, and Todd Zenger. This paper is also much improved due to feedback from editors and peer review.

Theory Is All You Need:
AI, Human Cognition, and Decision Making

ABSTRACT

Artificial intelligence (AI) now matches or outperforms human intelligence in an astonishing array of games, tests, and cognitive tasks that involve high-level reasoning and thinking. Many argue that humans should—or will soon—be replaced by AI in situations involving high-level cognition, judgment, and decision making. We disagree. In this paper we first trace the historical origins of the idea of AI and human cognition as forms of computation. We highlight problems with the analogy between computers and minds as input-output devices, using large language models as an example. Human cognition is better conceptualized as a form of theorizing rather than information processing, data-based prediction, or Bayesian updating. AI uses a frequency or probability-based approach to knowledge and is largely backward-looking and imitative, while human cognition is forward-looking and capable of generating genuine novelty. We argue that AI's data-based prediction is different from human theory-based causal logic. We introduce the idea of data-belief asymmetries to highlight the difference, using the example of "heavier-than-air flight" to illustrate our arguments. Theories provide a cognitive mechanism for humans to identify new data, a way of intervening in the world, experimenting, and problem solving. Throughout the article we discuss the implications of our argument for strategy and decision making under uncertainty.

Key words: cognition, artificial intelligence, information processing, prediction, decisions, strategy, theory-based view

INTRODUCTION

Artificial intelligence (AI) now matches or outperforms humans in any number of games, standardized tests, and cognitive tasks that involve high-level thinking and strategic reasoning. For example, AI engines can readily beat humans in chess, which for decades served as a key benchmark of AI capability (Bory, 2019; Simon, 1985). AI systems also now perform extremely well in complex games (like Diplomacy or Stratego) that involve sophisticated negotiation, complex interaction with others, alliances, deception, and understanding other players' intentions (e.g., Ananthaswamy, 2022). Current AI models also outperform over 90% of humans in various professional qualification exams, like the Bar exam in law and the CPA exam in accounting (Achiam et al., 2023). AI has also made radical strides in medical diagnosis, beating highly-trained medical professionals in diagnosing some illnesses (e.g., Zhou et al., 2023). These rapid advances have led some AI scholars to argue that even the most human of traits, like consciousness, will in principle soon be replicable by machines (e.g., Butlin et al., 2023; Goyal and Bengio, 2022). In all, AI is rapidly devising algorithms that "think humanly," "think rationally," "act humanly," and "act rationally" (Csaszar and Steinberger, 2022).

Given the astonishing progress of AI, Daniel Kahneman asks (and answers) the logical next question: "Will there be *anything* that is reserved for human beings? Frankly, I don't see any reason to set limits on what AI can do…And so it's very difficult to imagine that with *sufficient data* there will remain things that only humans can do…You should replace humans by algorithms whenever possible" (2018: 609-610, *emphasis added*).

Kahneman is not alone in this assessment. Davenport and Kirby argue that "we already know that analytics and algorithms are better at creating insights from data than most humans," and that "this human/machine performance gap will only increase" (2016: 29). Many scholars claim that AI is likely to outperform humans in most—if not all—forms of reasoning and decision making (e.g., Grace et al., 2024, Legg and Hutter, 2007; Morris et al., 2023). Some argue that strategic decision making might also be taken over by AI (Csaszar, Ketkar and Kim, 2024), or that even science itself will be "automated" (Zhu, and Horton, 2024; for related arguments, see Agrawal et al., 2024; Zhu and Griffiths, 2024). One of the pioneers of AI, Geoffrey Hinton, argues that large language models are sentient and intelligent, and that "digital intelligence" will inevitably surpass human "biological intelligence"—if it has not already done so (see Hinton, 2023; also see Bengio et al., 2023).

Compared to machines, the cognitive and computational limitations of humans obvious. Humans are biased and boundedly rational (for a review, see Chater et al., 2018; also see Kahneman, 2003; Kahneman, 2011). Humans are selective about what data they attend to and sample, and they are susceptible to confirmation and hundreds of other cognitive biases (nearly two hundred as of last count). In short, humans are "boundedly rational"—significantly hampered by their ability to compute and process information (Simon, 1955), particularly compared to computers (cf. Simon, 1990). And the very things that make humans boundedly rational and poor at decision making, are seemingly the very things that enable computers to perform well on cognitive tasks. The advantage of computers and AI is that they can handle vast amounts of data and process it quickly and in powerful ways.

In this paper we offer a contrarian view of AI relative to human cognition—including its implications for strategy and decision making under uncertainty. We first revisit the historical origins of the claim that equates computation with human cognition. AI builds on the idea that cognition is a generalized form of information processing, an "input-output device." To illustrate cognitive differences between humans and computers, we use the example of large language models versus human language learning. Building on these differences, we argue that human cognition in important instances operates in forward-looking fashion—from theories to data. We introduce the notion of "data-belief (a)symmetry" and the role this respectively plays in explaining AI and human cognition, using "heavier-than-air" flight as an extended example. Human cognition is forward-looking, necessitating data-belief asymmetries which are manifest in theories, as well as human causal reasoning and experimentation. Human cognition is driven by theory-based causal logic which is different from AI's emphasis on data-based prediction. Theories enable the generation of *new and contrarian* data, observations, and experimentation. We highlight the implications of these arguments for decision making under uncertainty, along with briefly highlighting opportunities for considering human-AI hybrid systems.[1]

[1] We need to briefly comment on the title of this paper—"theory is all you need." Our title echoes the title of the "attention is all you need" article that introduced the transformer architecture which (among other technologies) gave rise to recent progress in AI (Vaswani et al., 2017). But just as "attention" is not *all* an AI system or large language model needs, so theory of course is not *all* that humans need. In this article we simply emphasize that theory is a foundational—often unrecognized—aspect of human cognition, one that is not easily replicable by machines and AI. We emphasize the

AI = MIND: A REVIEW OF COGNITION AS COMPUTATION

Modeling the human mind—thinking, rationality, and cognition—has been the central aspiration and ambition behind AI from the 1940s to the present (McCulloch and Pitts, 1943; Turing 1948; also see Simon, 1955; Hinton, 1992; McCorduck, 2004; Perconti and Plebe, 2020). As put by the organizers of the first conference on AI—held at Dartmouth in 1956—their goal was to "proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy et al., 2007: 12). The commonalities between models of AI and human cognition are not just historical, but these linkages have only deepened in the intervening decades (for a review, see Sun, 2023; also see Laird et al., 2017). Computation also underlies many other models of cognition, including the concept of mental models (Johnson-Laird, 1983), the Bayesian brain, and predictive coding or processing (e.g., Friston and Kiebel, 2009; Hohwy, 2013, 2020). In fact, cognitive scientist Johnson-Laird goes so far as to argue that "any scientific theory of the mind has to treat it as an automaton" (1983: 477).

AI sees cognition as a general form of computation, specifically where "human thinking is wholly information-processing activity" (Feigenbaum, 1963: 249; also see Simon, 1980). This logic is also captured by computational neuroscientist David Marr who states that "most of the phenomena that are central to us as human beings—the mysteries of life and evolution, of perception and feeling and thought—are primarily phenomena of information processing" (1982: 4; cf. Hinton, 2023). Both mind and machine are a type of generalized input-output device, where inputs such as stimuli and cues ("data") are processed to yield varied types of outputs, including decisions, capabilities, behaviors, and actions (Simon, 1980; 1990; Hasson et al., 2020; McCelland and Rumelhart, 1981). This general model of information processing has been applied to any number of issues and problems at the nexus of AI and cognition, including perception, learning, memory, expertise, search, and decision making (cf. Russell and Norvig, 2022). Furthermore, the idea of human mental activity as computation is pervasive in evolutionary arguments. For example, Cosmides and Tooby focus on the "information-processing architecture of the human brain" and further argue that "the brain is a computer, that is, a physical system that was designed to process information" (2013: 202-203).

---

role of theory in human cognition, particularly the ways in which humans counterfactually think about, causally experiment, and practically "intervene" in the world.

The earliest attempts to develop machines that simulate human thought processes and reasoning focused on *general* problem solving. Newell and Simon's (1959) "general problem solver" (GPS) represented an ambitious effort to (try to) solve *any* problem that could be presented in logical form. GPS used means-ends analysis, a technique that compared a current state to the desired state (or goal), identified the differences, and then applied operators (actions) to reduce these differences. The early excitement associated with GPS and other AI models—and their ability to mimic human intelligence and thought—was pervasive. As put by Herbert Simon in 1958, "there are now in the world machines that think, that learn and create. Moreover, their ability to do these things is going to increase rapidly until—in a visible future—the range of problems they can handle will be coextensive with the range to which the human mind has been applied" (Simon and Newell, 1958: 8).

Early models like GPS provided the foundations for general cognitive architectures like SOAR and ACT-R (Anderson, 1990; Laird, Newell and Rosenbloom, 1987). The enthusiasm for these types of general models of cognition and AI continues to this day. Kotseruba and Tsotsos (2020) offer an extensive survey of over two hundred different "cognitive architectures" developed over the past decades. The ultimate goal of all this research into cognition, they argue, "is to model the human mind, eventually enabling us to build human-level artificial intelligence" (2020: 21). However, while various cognitive architectures related to AI hope to be general—and to mimic or even exceed human capability—their application domains have turned out to be extremely narrow and specific in terms of the problems they actually solve. But despite limited success in generalizing early models of AI (specifically, from the late 1950s to the 1990s), excitement about the possibility of computationally modeling human cognition did not wane. Simon's frequent collaborator, Alan Newell, argued that "psychology has arrived at the possibility of unified theories of cognition," specifically where "AI provides the theoretical infrastructure for the study of human cognition" (1990: 40). This unified approach builds on the premise that humans share certain "important psychological invariants" with computers and artificial systems (Simon, 1990: 3). This logic has also been captured by such ideas as "computational rationality" (Gershman et al., 2015).

In all, to this day there are ongoing calls for and efforts to develop so-called "common model of cognition"—or as put by others, a "standard model of the mind" based on AI (Laird et al., 2017; cf. Kralik et al., 2018). The call for general models has been born out of a frustration with the aforementioned proliferation of cognitive models that claim to be general, despite the fact that these models are highly heterogeneous, and any given model is highly focused on solving very specific tasks and problems. The effort to create a "meta"-model of cognitive AI—a model that different proponents of various cognitive architectures could agree on—has so far led to the identification of relatively generic elements. These models include basic elements like perception (focused on incoming stimuli or observations of the state of the world), different types of memory (and accompanied learning mechanisms), which in turn are linked to various motor systems and behaviors (Laird et al., 2017).

So far our short and informal review of the AI-cognition nexus has largely focused on symbolic systems, so-called "good old-fashioned AI."[2] Another approach to AI and modeling the human mind—called subsymbolic—also builds on the idea of information processing and computation but emphasizes bottom-up learning. These models also see the mind (or brain) as an input-output device. But the emphasis is on learning things "from scratch"—that is, learning directly from data. Vast inputs and raw data are fed to these systems to recognize correlations and statistical associations, or in short, patterns. The weakness of the aforementioned symbolic systems is that these approaches are only useful for relatively static contexts which do not meaningfully allow for any form of dynamic, bottom-up learning from data or environments.

The foundations of subsymbolic AI were laid by scholars seeking to understand the human brain, particularly perception. Rosenblatt (1958, 1962; building on Hebb, 1949) proposed one of the earliest forms of a neural network in his model of a "perceptron," which is the functional equivalent of an artificial neuron. Rosenblatt's work on the perceptron aimed to replicate the human neuron, which when coupled together would resemble human neural networks. Since modern artificial neural networks—including convolutional, recurrent, autoencoders, generative adversarial networks—build on this broad foundation (e.g., Aggarwal,

[2] These approaches are an attempt to develop intelligence by the manipulation of symbols—which represent objects, concepts, or states of the world—specifically through logical rules and the development of heuristics. The symbolic approach to cognitive AI models the world using symbols, and then uses logical operations to manipulate these symbols to solve problems. This represents a rule-based and top-down approach to intelligence. It is top-down in the sense that it starts with a high-level focus on understanding a particular problem domain and then breaking it down into smaller pieces (rules and heuristics) for solving a specific task. Perhaps the most significant applications in AI—between the 1950s and late 1980s—were based on these rule-based approaches. One of the more successful applications of an AI-related problem solver was the backward chaining expert system MYCIN, which was applied to the diagnosis of bacterial infections and the recommendation of appropriate antibiotics for treatment (Buchanan and Shortliffe, 1984). The goal of a system like this was to mimic the judgments of an expert decision maker. The model was a type of inference engine that used various pre-programmed rules and heuristics to enable diagnosis. In all, AI that is based on symbolic systems represents a top-down approach to computation and information processing that seeks to develop a rule- or heuristic-based approach to replicate how a human expert might come to a judgment or a decision.

2018; LeCun, Bengio and Hinton, 2015), it is worth briefly highlighting the general architecture of this approach. The architecture of the multi-layer perceptron includes layers that resemble the sensory units (input layer), association units (hidden layer), and response units (output layer) of the brain. This structure is very much the foundation of modern neural networks (Hinton, 1992; Rumelhart et al., 1986) and the basis for the radical advances made in areas such as AI image recognition and computer vision (Krizhevsky, Sutskever and Hinton, 2012).[3] The process of learning in a neural network—as specified by Rosenblatt—begins with stimuli hitting the sensory units, generating a binary response that is processed by the association cells based on a predetermined threshold. The association cells then send signals to the response area, which determine the perceptron's output based on the aggregated inputs from the association cells. The perceptron's learning mechanism is based on feedback signals between the response units and the association units, allowing the network to learn and self-organize through repeated exposure to stimuli. So-called Hebbian learning (Hebb, 1949)—which posits the relatively cliché but important idea that "neurons that fire together, wire together"—was the precursor to these types of feedback-based learning processes and many modern concepts of neural network theory.

In the intervening decades, research on artificial neural networks has progressed radically from simple classifiers to highly complex, multilayer non-linear models capable of sophisticated feature learning and pattern recognition through weighting and updates using large datasets (e.g., Aggarwal, 2018; Shazeer et al., 2017). Various forms and combinations of machine learning types—for example: supervised, unsupervised, and reinforcement learning—have enabled radical breakthroughs in image recognition and computer vision, recommender systems, game play, text generation, and so forth. And commensurate interest in the interaction between human neural networks and AI—various forms of learning—has continued within the cognitive sciences. This includes work on learning the structure of the environment (Friston et al., 2023; also see Hasson et al., 2020), meta learning (Lake and Baroni, 2023) or so-called "meta-learned models of cognition" (Binz et al, 2023), as well as inductive reasoning by humans and AI (Bhatia, 2023), and inferential learning (Dasgupta et al., 2020). Many of these models of learning build on neural networks in various forms, as well as related approaches.

[3] While these models emerged seemingly out of nowhere, it is important to understand that the foundations were laid decades ago (Buckner, 2023).

Now, our overall purpose is *not* to review the exhaustive details of these models and their underlying architectures, particularly as excellent reviews can be found elsewhere (e.g., Aggarwal, 2018; Russell and Norvig, 2022; Goodfellow et al., 2016). Rather, in the above we simply seek to offer a high-level overview of these models of AI—from the 1950s to the present—and their links to human cognition and mind. In all of this work, cognition and computation (and AI) are seen as deeply connected: the underlying premise of this work is that machines and humans are a form of input-output device, where the same underlying mechanisms of information processing and learning are at play. The focus on computation and information processing also is the axiomatic basis for the concept of bounded rationality (for a review, see Felin, Koenderink, and Krueger, 2017). Bounded rationality is focused on human "computational capacities" and their limits (Simon, 1955: 99)—and this idea has deeply shaped fields such as economics, decision theory, strategy, and the cognitive sciences (e.g., Chater et al., 2018; Gigerenzer and Goldstein, 2024; Kahneman, 2003; Puranam et al., 2015).

In all, we disagree with the idea that AI and human cognition share significant similarities as forms of computation, for reasons to be discussed next. That said, our aim in making this claim is *not* to take away from the exciting breakthroughs in AI. Rather, we highlight how the analogy between AI and humans quickly breaks down when it comes to understanding the mind and cognition, with important derivative consequences for how we think about judgment, strategy, and decision making under uncertainty. In the next section we delve into a specific example, namely language learning by machines versus humans, to enable us to make this point more carefully.

MACHINE VERSUS HUMAN LEARNING: DIFFERENT INPUTS, DIFFERENT OUTPUTS

While the input-output model of minds and machines—whether we are talking about symbolic or subsymbolic approaches—has been a central thesis of AI and cognitive science, next we highlight some important differences between machine learning and human learning. An apt context for highlighting these differences is to focus on language. Language arguably is "the most defining trait of human cognition (language and its relation with thought)" and therefore it "can be a true 'window into the mind' "(Chomsky, 2020: 321; also see Pinker, 1994).[4] Language provides an important "test" and context for understanding

[4] Recent comparisons between large language models (LLMs) and humans have revealed intriguing insights into their formal versus functional linguistic competence. In humans these two forms of competence rely on different neural mechanisms (Mahowald et al., 2024).

human and artificial intelligence. Furthermore, some have already argued that large language models are sentient, with a few even arguing that they already closely mirror or exceed human cognition (e.g., Binz and Schulz, 2023; Hinton, 2023)—an assumption which we challenge.

At the most basic level, to study any system and its behavior we need to understand its inputs and outputs. Alan Turing (1948) argued that any form of intelligence, whether human or machine, can be studied as an input-output system. In discussing the possibilities of artificial intelligence—or "intelligent machinery" as he called it—Turing made the analogy to an "untrained infant brain." An infant brain is largely a blank slate, "something like a notebook" with "*little* mechanism, and lots of *blank* sheets" (1950: 456, *emphasis added*; cf. Turing, 1948). According to Turing, these blank sheets are (or need to be) filled with inputs via the process of training and education. Through the early course of its life, an infant or child is taught and receives inputs in the form of language and spoken words that it hears and encounters. Education and training represent the inputs that eventually account for human linguistic capacities and outputs. And in the same way, Turing argues, one can think of an "analogous teaching process applied to machines" (1948: 107), where machines learn from their inputs. Turing lists various settings in which a thinking machine might show that it has learned—including games like chess or poker, cryptography, or mathematics—and he argues that "learning of languages would be the most impressive, since it is the most human of these activities" (Turing, 1948: 117). As human and machine learning are often seen as a similar process, we next focus on key differences using language learning as our example. We then highlight the implications of these differences in learning for decision making and knowledge generation both in scientific and economic contexts.

How Machines Learn Language

To illustrate the process of machine learning, next we carefully consider modern large language models (LLMs) and how they learn. LLMs offer a useful instantiation of machine learning. Learning is essentially generated from scratch—bottom up, directly from the data—through the introduction of vast amounts of training data and the algorithmic processing of the statistical associations and interactions amongst that data. In the context of an LLM, the training data is composed of enormous amounts of words and text, pulled together from various public sources and the Internet. To appreciate just how much data and training these models incorporate, the latest LLMs (as of early 2024) are estimated to include some 13 trillion tokens (a token being the rough equivalent of a word). To put this into context, if a human tried to read this text—say at a speed of 9,000 words/hour (150 words/minute)—it would take over 164,000 years to read the 13 trillion words of a training dataset.

The vast corpus of text used to train an LLM is tokenized to enable natural language processing. This typically involves converting words (or sub-word units or characters) into numerical sequences or vectors. To illustrate, a sentence like “The cat sat on the mat” might be tokenized into a sequence like [10, 123, 56, 21, 90, 78]. Each token is passed through an embedding layer, which converts the token into a dense vector representation that captures semantic information, such as its frequency and positional embedding. The embedding layer has its own set of parameters (weights) that are learned during training. The attention mechanism introduced with the “transformer” architecture (Vaswani et al, 2017), touched on by us previously, allows the model to consider each token in context of all other surrounding tokens, thus to gain an understanding of the wider context. Deep artificial neural networks have turned out to be extremely general and applicable not just to text but varied domains like image recognition and computer vision, including multi-modal applications that combine various types of data.[5]

From the vast data that serves as its training input, the LLM learns associations and correlations between various statistical and distributional elements of language: specific words relative to each other, their relationships, ordering, frequencies, and so forth. These statistical associations are based on the patterns of word usage, context, syntax, and semantics found within the training dataset. The model develops an “understanding” of how words and phrases tend to co-occur in varied contexts. The model does not just learn associations but also understands correlations between different linguistic elements. In other words, it discerns that certain words are more likely to appear in specific contexts.

Now, while the above is not a technical introduction to LLMs, it offers the broad outlines of the process to the degree that it is relevant for our argument (for a detailed review, see Chang et al., 2024; Minaee et al., 2024; Naveed et al., 2023; also see Resnik, 2024). The end-result of this training is an AI model that is

[5] In terms of the training of an LLM, the tokenized words are submitted for algorithmic processing based on a predetermined sequence or input length. Sequence length is important because it allows the LLM to understand context. The (tokenized) text is not fed into the system as one long string but rather in chunks of predetermined length. This predetermined length is variously called the context window, input or sequence length, or token limit. Recent LLM models (as of early 2024) typically use input lengths of 2,048 tokens. (Newer models are exploring longer sequence lengths.) Therefore, a 13 trillion token training dataset is parsed into 2,048-length sequences, enabling the algorithm to learn language. Learning language is a statistical exercise where the LLM learns from the word patterns, context, and dependencies found in the training data. It then uses this learning to stochastically generate outputs through next-word prediction.

capable of language: more specifically, the model is capable of generating fluent and coherent text by using a stochastic approach of "next-word prediction" in response to a prompt. Based on this broad outline of how an LLM is trained, we compare this to how humans learn language. We should reiterate, as discussed at the outset of this article, that the basic premise behind models of AI is that there is a symmetry between how machines and humans learn. We think it is important to carefully point out differences, as these provide the foundation for our subsequent arguments about cognition and decision making.

How Humans Learn Language, Compared to Machines

The differences between human and machine learning—when it comes to language (as well as other domains)—are stark. While LLMs are introduced to and trained with trillions of words of text, human language "training" happens at a much slower rate. To illustrate, a human infant or child hears—from parents, teachers, siblings, friends, and their surroundings—an average of roughly 20,000 words a day (e.g., Gilkerson et al., 2017; Hart and Risley, 2003). So, in its first five years a child might be exposed to—or "trained" with—some 36.5 million words. By comparison, LLMs are trained with trillions of tokens within a short time interval of weeks or months.[6]

The inputs differ radically in terms of quantity (sheer amount), but also in terms of their quality.[7] Namely, the *spoken* language that an infant or young child is (largely) exposed to is different from the written language on which an LLM is trained on. Spoken language differs significantly from written language in terms of its nature, structure, and purpose. Here the research on the differences between spoken and written language is highly instructive (e.g., Biber, 1991). Spoken language is spontaneous (not meaningfully edited), informal, repetitive, and often ephemeral. Written language—on the other hand—is visual and permanent, more carefully crafted, planned, and edited. It is also denser, featuring more complex vocabulary (e.g., Halliday, 1989; Tannen, 2007). More importantly, the functional purposes and uses of spoken versus written

[6] For an infant to be exposed to the same 13 trillion tokens represented by the training of current LLMs, it would take or roughly 1.8 million years.

[7] Of course, an infant is not just "trained" through the language it might be exposed to by auditory means, but also through other modalities and systems (including visual, olfactory, gustatory, and tactile ones). LLMs are largely monomodal, though various multimodal models of AI are of course in development. But setting aside questions of multimodality or even the "amount" of text or information that a system might be trained with, there are also deeper questions. That is, *how* humans are able to learn from the things they encounter in the first place, and *what* they learn (or what humans notice in the first place), is a key puzzle. Undoubtedly the biological nature and evolutionary history of humans is central to understanding these types of questions, as is the associated ability of humans—as we emphasize in this paper—to engage with their surroundings in novel and forward-looking ways.

language differ significantly. Spoken language is immediate, interactive, focused on coordinating, expressing, and practically doing things. While written language also serves these purposes, the emphasis is more on the communication of complex information. The vast bulk of the training data of the LLM is not conversational (for models trained on spoken language or "raw audio," see Lakhotia et al., 2022). Rather, written language is more carefully thought-out. An LLM is likely to be trained with the works of Shakespeare and Plato, academic publications, public domain books (e.g., from Project Gutenberg), lyrics, blog posts, news articles, and so forth. This data is far "cleaner," far more correct grammatically, and organized. Arguably the inputs received by an LLM—in the form of written, edited and published text—are linguistically far superior. In a statistical sense, LLM training data contains less "noise" and thus offers greater predictive power. Even the vast stores of Wikipedia articles that are included in most LLM training datasets are the end result of thousands of edits to ensure readability, accuracy, and flow.

Clearly humans learn language under different conditions and via different types of inputs. In short, it can readily be argued that the human capacity for language develops differently from how machines learn language in both quantity and quality. Humans (somehow) learn language from extremely sparse, impoverished, and highly unsystematic inputs and data (Chomsky, 1975). Compared to LLMs, human linguistic capabilities are radically "underdetermined" by the inputs. That is, the relatively sparse linguistic inputs can scarcely account for the radically novel outputs generated by humans.[8]

Beyond the quantitative and qualitative differences in inputs (when it comes language learning by LLMs versus humans), it is important to compare the linguistic outputs and capabilities of machines versus humans. In terms of output, LLMs are said to be "generative" (the acronym GPT stands for "generative pretrained transformer"). But in what sense are LLMs generative? They are generative in the specific sense that they are able to create novel outputs by probabilistically sampling from the vast combinatorial possibilities

---

[8] This logic is aptly captured by Chomsky: "One can describe the child's acquisition of knowledge of language as a kind of theory construction. Presented with *highly restricted data*, he constructs a theory of language of which this *data is a sample* (and, in fact, a *highly degenerate sample*, in the sense that much of it must be excluded as irrelevant and incorrect—thus the child learns rules of grammar that identify much of what he has heard as *ill-formed, inaccurate, and inappropriate*). The child's ultimate knowledge of language obviously *extends far beyond the data presented to him*. In other words, the theory he has in some way developed has a predictive scope of which the data on which it is based *constitute a negligible part*. The normal use of language characteristically involves new sentences, sentences that bear no point-by-point resemblance or analogy to those in the child's experience (1975: 179, *emphasis added*)." Our goal is not to endorse Chomsky's theory of universal grammar. Rather, we concur with this specific quote in terms of its characterization of the input-output relationship, where human linguistic outputs are *under*determined by the inputs children receive. Broadly this also links to the alternative approach that we focus on (the theory-based view of cognition), discussed in the second half of the paper.

in the associational and correlational network of word frequencies, positional encodings, and co-occurrences encountered in the training data (Vaswani, 2017).[9] The LLM is generative in the sense that the text that is produced is not simply plagiarized or copied verbatim from existing sources contained in the pretraining data (McCoy et al., 2023). In the process of generating text, the parameters (weights and biases) determine how much influence different parts of the training data probabilistically have on the output. For example, in a sentence completion task, the weights—developed from the corpus of the training data—help the model decide which words are most likely to come next, based on the context provided by the input. The output is statistically derived from the training data's underlying linguistic structure. The outputs therefore have compositional novelty (in terms of novel ways of saying the same thing—more on this below), and they also manifest some analogical generalization (McCoy et al., 2023). That said, any assessment of how "good" an LLM is needs to recognize "the problem that [LLMs] were trained to solve: next-word prediction" (McCoy et al., 2023). And as next-word prediction engines, LLMs certainly demonstrate exceptional capabilities.

BEYOND MIRRORING: CAN AI GENERATE GENUINE NOVELTY?

So far we have summarized the central elements of a particular AI system—an LLM—and compared it with humans. Next we further address whether an AI can be said to be "intelligent" and whether it can make forward-looking decisions and be used to generate genuine novelty. While our focus remains on LLMs, we extend our arguments to other forms of AI and cognitive approaches that focus on data and prediction. We concurrently raise central implications of these arguments to the question of decision making under uncertainty.

AI: Intelligence and New Knowledge?

As we have foreshadowed above, an AI like an LLM seems to "mirror" the inputs it has been trained with rather than meaningfully manifest some form of intelligence. But beyond next-word prediction and linguistic fluency, could an LLM do a better job than humans in decision making under uncertainty (e.g.,

[9] Relative to the idea of next-word prediction (and the "draw" of the next word), there are different ways for this to happen. For example, a model might always pick the most likely next word (greedy). Or a model might explore multiple sequences simultaneously (beam search), along with many other approaches (top-k sampling, top-p sampling etc). In practice, different types of prompts (depending on prompt context, length, tone, style) lead to different types of sampling and next-word prediction (Holtzman et al., 2019), as will changing the "temperature" setting of the model.

Csaszar et al., 2024; cf. Kahneman, 2018)—or could an LLM perhaps even "automate" science and scientific reasoning itself (e.g., Manning et al., 2024)?[10]

Certainly LLMs seem to manifest sparks of intelligence. But intelligence is not simply memorization, the ability to restate or paraphrase information in various ways. LLMs appear intelligent by capitalizing on the *fact that the same thing can be stated, said, and represented in indefinite ways*. This is readily illustrated by the fact that the revolutionary breakthrough that gave rise to LLMs—the transformer architecture—was developed in the context of language translation (Vaswani et al., 2017). Accordingly, LLMs can be seen as translation generalized. LLMs can be seen as generalized technology for *translating one way of saying things into another way of saying the same thing*. Translation after all is an effort to represent and accurately mirror something in a different way—to represent the same thing in a different language or with a different set of words, or more abstractly: to represent the same thing in a different format. LLMs serve this representational and mirroring function remarkably well. This representational and mirroring function from language to language is generalized to a process that takes one way of saying something and generates another way of saying the same thing. Stochastic next-word prediction—using the weights and parameters found in vast training datasets and probabilistically drawing from this training—allows for surprisingly rich combinatorial outputs. The learning of the LLM is embodied in the relationships found between words which are sampled to enable stochastic generativity, where the outputs mirror past inputs. But, as we will discuss, the fluency with which LLMs seem to generate outputs dupes us into seeing them as intelligent—as if they are engaging in far more than mere translation. With vast data, an LLM is good at fluently predicting the next word.

Before revisiting our question of whether an AI like an LLM could actually engage in some form of forward-looking decision making, it is worth highlighting metaphorical similarities between AI and cognitive architectures based on prediction. For example, consider a cognitive approach like predictive processing (Pezzulo, Parr and Friston, 2024: a rough synonym for active inference, the free energy principle, the Bayesian

[10] AI can, of course, be (and has been) a powerful aid in scientific discovery. For example, modern AI techniques have analyzed astronomical datasets far more quickly and accurately than humans, helping identify new planets and celestial phenomena, as seen with Kepler's laws of planetary motion. Similarly, DeepMind's AlphaFold has revolutionized protein structure prediction, a critical task for understanding biological processes and developing new medications (e.g., Jumper et al., 2021). Yet it is important to state that in both of these cases AI is not somehow independently doing the science by forming hypotheses or conducting experiments, but that these hypotheses were provided by human scientists in the form of patterns and reward functions, respectively. AI has significantly accelerated research by enabling scientist to process large datasets and uncover novel patterns, allowing scientists to focus on hypothesis generation and experimental design rather than "number crunching."

brain, and predictive coding). At a high level, both LLMs and predictive processing seek to engage in a similar process, namely, in error minimization and iterative optimization, where the systems are essentially navigating a high-dimensional space to find a state that minimizes both error and surprise. LLMs learn from the training data and predictive processing learns from its environment (cf. Hohwy, 2020). LLMs aim to reduce the difference between their predictions (the next word in a sentence) and the actual outcomes (the real next word), thereby improving their accuracy. Predictive processing, as a cognitive theory, posits that the brain continuously predicts sensory input and minimizes the error between its predictions and actual sensory input. The capability of each to predict—whether a word or a perception—is a function of past inputs. Large language models seek to predict the most likely next word—based on training data—and active inference seeks to predict the most likely next percept or action. Both approaches are seeking to reduce surprise—or prediction as "error minimization" (Hohwy, 2013).[11] Back-propagation, a fundamental mechanism in training neural networks, and the concept of error minimization in predictive processing (Friston et al., 2009) share a broad conceptual similarity in that both involve iterative adjustments to minimize some form of error or discrepancy. Both create a prediction based on past inputs. Both back-propagation and error minimization in predictive processing involve adjusting an internal model (neural network weights in AI, and hierarchical brain models in neuroscience) to reduce error (or, in machine learning terms, minimize the loss function).

With this architecture—focused on error minimization and surprise reduction—can an LLM or any prediction-oriented cognitive AI truly generate some form of new knowledge? Beyond memorizing, translating, restating, or mirroring the text with which it has been trained, can an LLM generate new knowledge?

We do not believe LLMs or input-output based cognitive systems can do this, at least not beyond random flukes that might emerge due to their stochastic nature.[12] There is no forward-looking mechanism or unique causal logic built into these systems. It is important to clearly delineate *why* this is the case, as some argue and anticipate that LLMs will replace human decision makers in uncertain contexts like strategy and even science itself. For example, Csaszar et al (2024) argue that "the corpora used to train LLMs encompass

[11] This leads to the problem of surprise and the famous "dark room" problem of predictive processing. For an attempt to deal with this, see Clark, 2018.
[12] Though we of course recognize that there is significant disagreement on this point (for example, related to AI versus human creativity, see Franceschelli and Musolesi, 2023).

information necessary for [strategic decision making], including consumer preferences, competitor information, and strategy knowledge" and point to how an AI can use various decision making tools to generate business plans and strategy (Csaszar et al., 2024). And Manning et al (2024) even argue that LLMs will "automate" social science given their seeming ability to *generate* hypotheses and causal models, including testing them.

These claims are vastly overstated. One way to think about this is that a prediction-oriented AI like an LLM can essentially be said to possess "Wikipedia-level" knowledge. On any number of topics (if contained in the training data), an LLM can summarize, represent, and mirror the words they have encountered in various different and new ways. On any given topic—again, if sufficiently represented in the training data—an LLM could generate indefinite numbers of coherent, fluent, and well-written Wikipedia articles. But just as a subject-matter expert is unlikely to learn anything new about their specialty from a Wikipedia article within their domain of expertise, so an LLM is unlikely to somehow bootstrap knowledge beyond the combinatorial possibilities of the word associations it has encountered in the past.

There is also good evidence suggests that when an LLM encounters (is prompted with) a reasoning task, it merely *reproduces* the linguistic answers (about reasoning) it has encountered in the training data rather than engaging in any form of actual, on-the-fly reasoning. If the wording of a reasoning task—like the Wason selection task or the Monty Hall problem—is changed only slightly, LLM performance declines significantly below human performance, where the mistakes of the LLM are glaringly obvious to humans (e.g., Hong et al., 2024; Macmillan and Musolesi, 2024). LLMs are not meaningfully engaged in any form of real-time reasoning themselves (as seemingly assumed by Manning et al., 2024 and many others), rather they are merely repeating the word structures associated with reasoning, which they have encountered in the training data. Put differently, LLMs memorize the words associated with reasoning (in highly specific tasks) but are not engaged in reasoning on-the-fly.[13] This is why Francois Chollet (2019) has created the "abstraction and reasoning corpus"—as a challenge or test to see if an AI system can actually solve *new* problems (ones it has not

[13] The capabilities of AI are of course rapidly evolving and future developments are hard to anticipate. In this paper we have discuss AI in its *past and current state*—comparing it with human cognition—rather than speculate about what AI might be capable of in the future. It might be that the forms of human reasoning and cognition that we emphasize (and claim, in this paper, to be unique to humans) could be mimicked or replicated by future AI systems.

encountered in its training data), without merely resorting to memorized answers and solutions encountered in the past (which captures the present state of AI systems including LLMs).[14]

That said, our goal is not to dismiss the remarkable feats of LLMs, nor other forms of AI or applications of machine learning. The fact that an LLM can outperform most humans in varied types of tests and exams is remarkable (Achiam et al., 2023). But this is because it has encountered this information and memorized it. An LLM has a superhuman capacity for memorization, and an ability to summarize memorized word structures in diverse, fluent, and novel ways. In all, certainly the idea of LLMs as "stochastic parrots" or "glorified auto-complete" (Bender et al., 2021) underestimates their ability. But equally, ascribing LLMs the ability to actually reason and generate new knowledge vastly overestimates their ability.

We might concur that LLMs are powerful and creative "imitation engines" (in stochastically assembling words), though not linguistically innovative compared to children (see Yiu, Kosoy, and Gopnik, 2023). The idea that LLMs somehow generate new-to-the-world knowledge—or feature something like human consciousness—seems to be a significant stretch (though, see Butlin et al., 2023; Hinton, 2023). In sum, the generativity of these models is a type of "lower-case-g" generativity that shows up in the form of the unique sentences that creatively summarize and repackage existing knowledge.

To illustrate the problem of generating new knowledge with an LLM, imagine the following thought experiment. Imagine an LLM in the year 1633, where the LLM's training data incorporated all the scientific and other texts published by humans to that point in history. If the LLM were asked about Galileo's heliocentric view, how would it respond? Since the LLM would probabilistically sample from the association and correlation-based word structure of its vast training data—again, everything that has so far been written (including all the scientific writings about the structure of the cosmos)—it would only restate, represent, and mirror the accumulated scientific consensus. The training dataset would *overwhelmingly* feature texts with word structures supporting a geocentric view, in the form of the work of Aristotle, Ptolemy, and many others. Ptolemy's careful trigonometric and geometric calculations, along with his astronomic observations, would be included in support of a geocentric view, as represented in the many texts that would have summarized the

[14] Beyond the ability of a human or AI to solve previously-unseen, *new* problems (which is the focus of Chollet's ARC challenge), there is an even higher form of intelligence in being able to specify and formulate problems in the first place (Felin and Zenger, 2017). This is a skill that is manifest in humans—in theorizing and causal reasoning—but not evident in AI. As we discuss later, it was the ability of the Wright brothers to formulate the right problems (lift, propulsion, and steering) that enabled them to then identify the right data, specific forms of experimentation and relevant solutions.

geocentric view (like de Sacrobosco's popular textbook *De saphera mundi*). These texts would feature word associations that highlight how the motions and movements of the planets could be predicted with remarkable accuracy with the predominant geocentric view. The evidence—as inferred from the repeated word associations found in the training data—would overwhelmingly be against Galileo. LLMs do not have any way of accessing truth (for example through experimentation or counterfactuals) beyond mirroring and restating what is found in the text.

Even if alternative or heretical views were included in the training data (like the work of Copernicus, even though his work was largely banned), the logic of this work would be dwarfed by all the texts and materials that supported the predominant geocentric paradigm.[15] The overwhelming corpus of thousands of years of geocentric texts would vastly outweigh Galileo's view—or anything supporting it. *An LLM's model of truth or knowledge is statistical, relying on frequency of probability.* Outputs are influenced by the frequency with which an idea is mentioned in the training data, as reflected by associated word structures. For example, the frequency with which the geocentric view has been mentioned, summarized, and discussed in the training data necessarily imprints itself onto the output of the LLM as truth. As the LLM has no actual grounding in truth—beyond the statistical relationships between words—it would say that Galileo's view and belief is delusional and in no way grounded in science.

A neural network like an LLM might in fact include any number of delusional beliefs, including beliefs that turned out to *eventually* be correct (like Galileo's), but also beliefs that *objectively* were (and still are) delusional. Ex ante there is no way for an LLM to arbitrate between the two. For example, the eminent astronomer Tyco Brahe made and famously published extensive claims about astrology, the idea that celestial bodies and their movement directly impact individual human fates as well as political and other affairs. His astrological writings were popular not just among some scientists, but also among the educated elite. A hypothetical LLM (in 1633) would have no way of arbitrating between Galileo's (seeming) delusions about heliocentrism nor Brahe's (actual) delusions about astrology. The LLM would argue against heliocentrism and for astrology. The LLM can only represent and mirror the predominant and existing conceptions—in this

[15] While Copernicus's *On the Revolution of the Heavenly Spheres* was published in 1543, the theory contained within the book represented a fringe view within science. Given the fringe nature of the Copernican view, his book was withdrawn from circulation and eventually censored (Gingerich and Maclachlan, 2005).

case, the support for the geocentric view of the universe—it finds in the frequencies and statistical association of words in its training data.

In sum, it is important to recognize that the way an LLM gets at truth and knowledge is via a statistical exercise of finding *more frequent* mentions of a true claim (in the form of statistical associations between words) and less frequent mentions of a false claim. LLM outputs are stochastically or probabilistically drawn from the statistical associations of words in the training data. When an LLM makes truthful claims, these are an epiphenomenon of the fact that true claims happened to have been made more frequently in the training data. There is no other way for the LLM to assess truth or reason. Truth emerges as a byproduct of statistical patterns and frequencies rather than from the LLM developing an intrinsic understanding of—or ability to bootstrap or reason—what is true or false in reality.

Some LLMs have sought to engineer around the problem of their frequency-based and probabilistic approach by creating so-called "mixture of experts" models where the outputs are not simply the "average" result of "outrageously" large neural networks, but can be fine-tuned toward some forms of expertise (Du et al., 2023; Shazeer et al., 2017). Another approach is retrieval-augmented generation, which uses the general linguistic abilities of the LLM but limits the data used for prediction to a confined and pre-selected set of sources (Lewis et al., 2020). Furthermore, ensemble approaches—which combine or aggregate diverse architectures or outputs—have also been developed (Friedman and Popescu, 2008; Russell and Norvig, 2022). However, even here the outputs would necessarily also be reflective of what any particular experts have said within the specified data, rather than any form of forward-looking projection or on-the-fly causal reasoning on the part of the LLM. This problem is further compounded in situations that are characterized by high levels of uncertainty and novelty (like many forms of decision making), where the idea of expertise or even bounded rationality is hard to specify given an evolving and changing world (Felin, Kauffman, Koppl and Longo, 2014).

Finally, it is critically important to keep in mind that the inputs of any LLM are *past* human inputs, and therefore outputs also roughly represent what we know so far. Inherently an LLM cannot go beyond the realms covered by the inputs. There is no mechanism to somehow bootstrap forward-looking beliefs about the future—nor causal logic or knowledge—beyond what can be inferred from the existing statistical associations and correlations found in the words in the training data.

The Primacy of Data versus Data-Belief Asymmetry

The central problem we have highlighted, so far, is that learning by machines and AI is necessarily backward-looking and imitative. Again, this should not be read as a critique of these models, rather, merely as a description of their structural limits. While they are useful for many things, an AI model—like an LLM—is not able to generate new knowledge or solve new problems. An LLM does not reason. And an LLM has no way of postulating beyond what it has encountered in its training data. Next we extend this problem to the more general emphasis on the primacy of data within both AI and cognitive science. Data itself of course is not the problem. Rather, the problem is that data is used in theory-independent fashion (cf. Anderson, 2007). To assure the reader that we are not caricaturing existing AI-linked models of cognition by simply focusing on LLMs, we also extend our arguments into other forms of cognitive AI.

The general emphasis on minds and machines as input-output devices places a primary emphasis on data. This suggests a model where data—such as cues, stimuli, text, images—essentially are "read," learned and represented by a system, whether it is human or computational one. The world (any large corpus of images, text, or the environment) has a particular statistical and physical structure, and the goal of a system is to accurately learn from it and reflect it. This is said to be the very basis of intelligence. As put by Poldrack, "any system that is going to behave intelligently in the world must contain *representations that reflect the structure of the world*" (2021: 1307, *emphasis added*; cf. Yin, 2020). Neural network-based approaches and machine learning—with their emphasis on bottom-up representation—offer the perfect mechanism for doing this, because they can "learn directly from data" (Lansdell and Kording, 2019; also see Baker et al., 2022). Learning is data-driven.[16] Of course, systems may not be able to learn perfectly, but an agent or machine can "repeatedly interact with the environment" to make inferences about its nature and structure (Binz et al., 2023). This is the basis of "probabilistic models of behavior" which view "human behavior in complex environments as solving a statistical inference problem" (Tervo, Tenenbaum, and Gershman, 2021).[17]

[16] The problems with this approach have not just been discussed by us. For example, Yin, 2020 for related points in the field of neuroscience.

[17] In the context of machine learning, it is interesting that while the approach is said to be "theory-free" (to learn directly from data), nonetheless the architects of these machine learning systems are making any number of top-down decisions about the design and architecture of the algorithms, and *how* the learning occurs and the types of outputs that are valued. These decisions all imply mini-theories of what is important—a point that is not often recognized (cf. Rudin, 2019). This involves obvious things like choice of data, model architecture, and hyperparameter settings, but also loss functions and metrics, regularization and generalization techniques, valued outputs, type of human reinforcement, and so forth.

Bayesian cognition also posits that learning by humans and machines can be understood in terms of probabilistic reasoning about an environment, as captured in Bayesian statistical methods (e.g., Griffiths et al., 2010). This framework conceptualizes sensory inputs, perceptions, and experiential evidence as data, which are continuously acquired from the environment and then used to update one's model of the world (or of a particular hypothesis). The cognitive process involves sampling from a probability distribution of possible states or outcomes, informed by incoming data. Crucially, Bayesian and related approaches to cognition emphasize the dynamic updating of beliefs—where prior knowledge (priors) is integrated with new evidence to revise beliefs (posterior), in a process mathematically described by the Bayesian formula (Pinker, 2021). This iterative updating, reflecting a continual learning process, acknowledges and quantifies uncertainty, framing understanding and decision making as inherently probabilistic. This probabilistic architecture is (very broadly) also the basis of large swaths of AI and the cognitive sciences.

It is worth reflecting on the epistemic stance, or underlying theory of knowledge, that is presumed here. Knowledge is traditionally defined as *justified* belief—and a belief is justified by data and evidence. As suggested by Bayesian models, we believe or know things to the extent to which we have data and evidence for them (Pinker, 2021). Beliefs should be proportionate to the evidence at hand, because agents are better off if they have an *accurate* representation or conception of their environment and the world (e.g., Schwöbel et al., 2018).[18] Knowledge can be seen as the accumulated inputs, data and evidence that make up our beliefs. And *the strength* or *degree* of any belief should be symmetrical with the amount of supporting data, or put differently, the *weight* of the evidence (Pinker, 2021; also see Dasgupta et al., 2020; Griffin and Tversky, 1992; Kvam and Pleskac, 2016). This is the basis of probabilistic models of cognition. These approaches focus on "reverse-engineering the mind"—from inputs to outputs—and "forges strong connections with the latest ideas from computer science, machine learning, and statistics" (Griffiths et al., 2010: 363). Overall, this represents a relatively widely-agreed upon epistemic stance, which also matches an input-output-oriented "computational theory of mind" (e.g., Rescorla, 2015) where humans or machines learn "through repeated interactions with an environment"—without "requiring any a priori specifications" (Binz et al., 2023). One way to summarize the above literature is that there needs to be a symmetry between one's belief and the corroborating data. A

[18] The predictive processing and active inference approach has many of these features (e.g., Parr and Friston, 2017)

rational decision maker will form (and weight) their beliefs about any given thing by taking into account the available data and evidence (Pinker, 2021).

But what about "edge cases?" That is, what about situations where an agent correctly takes in all the data and evidence yet somehow turns out to be wrong? Models based on rational information processing do not offer a mechanism for explaining change or new knowledge, nor an explanation of situations where data and evidence-based reasoning might lead to poor outcomes (cf. Felin and Koenderink, 2022). While learning-based models of knowledge enable "belief updating"—based on new data—*there is no mechanism for explaining where new data comes from*, or what data should be considered as relevant and what data should be ignored. And what if the data and evidence are contested? This is a particularly significant problem in contexts that feature rampant uncertainty, including any type of forward-looking decision making and scientific reasoning.

These situations are highly problematic for computational, input-output models of cognition that assume what we call *data-belief symmetry*. The basis of knowledge is the quest for truth (Pinker, 2021), which is focused on existing evidence and data. But we argue that data-belief *a*symmetry in fact is essential for the generation of new knowledge and associated decision making. The existing literature in the cognitive sciences has focused on one side of the data-belief asymmetry, namely its *downside*—the negative aspects of data-belief asymmetry (e.g., Kunda, 1990; Scheffer et al., 2022). This downside includes all the ways in which humans persist in believing something *despite* seemingly clear evidence to the contrary (Pinker, 2021). This includes a large literature which focuses on human biases in information processing—the suboptimal and biased ways that humans process, perceive, and use data and fail to appropriately update their beliefs. This is evident in the vast literatures that focus on various data-related pathologies and biases, including confirmation bias, selective perception and sampling, and availability bias. The emphasis on erroneous beliefs and human bias has powerfully influenced how we think about human nature and decision making within various social and economic domains (e.g., Bordalo et al., 2024; Chater, 2018; Gennaioli and Shleifer, 2018; Kahneman, 2011; Kahneman et al., 2021).

But what about the positive side of data-belief asymmetry? What about situations where beliefs appear delusional and distorted—seemingly contrary to established evidence and facts—but where these beliefs nonetheless turn out to be correct? Here we are specifically talking about beliefs that may outstrip, ignore, and go beyond existing evidence. Forward-looking contrarian views are essential for the generation of knowledge. Due to the statistical nature of AI-based computational systems (focused on correlations, associations, and averages from past data), they are not able to project forward in contrarian ways, given the implicit insistence on symmetry between data and beliefs. That said, notice that—as we will discuss—our focus on data-belief asymmetries is *not* somehow data-independent or untethered from reality. Rather, this form of data-belief asymmetry is forward-looking, where beliefs enable the identification of new data and experimental interventions, and the eventual verification of beliefs that previously were seen as the basis of distortion or delusion.

To offer a practical and vivid illustration of how data-belief symmetry can be problematic, consider the beliefs that were held about the plausibility of "heavier-than-air" human, powered, and controlled flight in the late 1800s and the early 1900s. (We introduce this example here and revisit it throughout the remainder of the manuscript.) To form a belief about the possibility of human powered flight—or to even assign it a probability—we would first want to look at the existing data and evidence. So, what was the evidence for the plausibility of human powered flight at the time? The most obvious datapoint at the time was that human powered flight was not a reality. This alone, of course, would not negate the possibility. So, one might want to look at all the data related to human flight attempts to assess its plausibility. Here we would find that humans have tried to build flying machines for centuries, and flight-related trials had in fact radically accelerated during the 19th century. All of these trials of flight could be seen as the data and evidence we should use to update our beliefs about the *implausibility* of flight. All of the evidence clearly suggested that a belief in human powered flight was delusional. A delusion can readily be defined as having a belief contrary to evidence and reality (Pinker, 2021; Scheffer, 2022): a belief that does not align with accepted facts. In fact, the DSM-4/5—the authoritative manual for mental disorders—defines delusions as "false beliefs due to incorrect inference about external reality" or "fixed beliefs that are not amenable to change in light of conflicting evidence."

Notice that many people at the time—naively, it was thought—pointed to birds as evidence for the belief that humans might also fly. This was a common argument.[19] But the idea that bird flight somehow provided hope and evidence for the plausibility of human flight was seen as delusional by scientists and put to

[19] As captured by a prominent engineer at the time: "There probably can be found no better example of the speculative tendency carrying man to the verge of the chimerical than in his attempts to imitate the birds, or no field where so much inventive seed has been sown with so little return as in the attempts of man to fly successfully through the air" (Melville, 1901: 820).

bed by the prominent scientist Joseph LeConte. He argued that flight was "impossible, *in spite of the testimony of birds*" (1888: 69). Like a good scientist and Bayesian, LeConte appealed to the data to support his claim. He looked at bird species—those that fly and those that do not—and concluded "there is a limit of size and weight of a flying animal." Weight was the critical determinant of flight. With his data, LeConte's pointed out that clearly no bird above the weight of 50 pounds is able to fly, and thus concluded that therefore humans cannot fly. After all, large birds like ostriches and emus are flightless. And even the largest flying birds—like turkeys and bustards—"rise with difficulty" and "are evidently near the limit" (LeConte, 1888; 69-76). Flight and weight are correlated. To this, Simon Newcomb—one of the foremost astronomers and mathematicians of his time—added that "the most numerous fliers are little insects, and the rising series stops with the condor, which, though having much less weight than a man, is said to fly with difficulty when gorged with food" (1901: 435).

The emphasis that prominent scientists placed on the weight of birds to disprove the possibility of human powered flight highlights one of the problems with data. It is hard to know what data and evidence might be *relevant* for a given belief or hypothesis. The problem is—as succinctly put by Polanyi—that "things are not labeled evidence in nature" (1957: 31). What is the relevant data and evidence in this context? Did flight have something to do with weight, size, or with other features like wings? Did it have something to do with the "flapping" of wings (as Jacob Degen hypothesized)? Or did it have something to do with wing shape, wing size, or wing weight?[20] Perhaps feathers were critical to flight. In short, it is hard to know what data might be relevant and useful.

Of course, not all our beliefs are fully justified in terms of direct empirical data that we ourselves have verified. We cannot—nor would we want to—directly verify all the data and observations that underlie our beliefs and knowledge. More often than not, for our evidence we rightly rely on the expertise, beliefs, or scientific arguments of others, which serve as "testimony" for the beliefs that we hold (Coady, 1992; Goldman, 1999). The cognitive sciences have also begun to emphasize this point. Bayesian and other probabilistic models of cognition have introduced the idea of "the reliability of the source" when considering

[20] Even if LeConte happened to be right that the delimiting factor for flight was weight (which of course is not the case), he also did not take into account—or more likely, was not aware of—findings related to prehistoric fossils. For example, in the mid and late 1800s, scientific journals reported about the discovery of prehistoric fossils—pelagornis sandersi—with wingspans of up to 20-24 feet and conjectures of a weight up to 130 pounds.

what data or evidence to use to update beliefs and knowledge (e.g., Hahn, Merdes, and Sydow, 2018; Merdes et al., 2021). This approach recognizes that not all data and evidence is equal. Who says what matters. For example, scientific consensus and expertise are an important source of beliefs and knowledge.

This is readily illustrated by our discussion of heavier-than-air flight. So, what might happen if we weight our beliefs about the plausibility of human flight by focusing on reliable, scientific sources and consensus? In most instances, this is a rational strategy. However, updating our belief on this basis when it comes to heavier-than-air flight would further reinforce the conclusion that human powered flight was delusional and impossible. Again, scientists like LeConte and Newcomb argued that flight was impossible by pointing to seemingly conclusive data and evidence. And not only should we base our belief on the evidence, but we should also weight the belief by the fact that prominent scientists were making these claims. LeConte for example became the eventual President of the leading scientific association at the time, the American Association for the Advancement of Science. And LeConte was scarcely alone. He was part of a much broader scientific consensus that insisted on the impossibility of human powered flight. For example, Lord Kelvin emphatically argued—while serving as President of the British Royal Society—that "heavier-than-air flying machines are impossible." This is ironic, as Kelvin's scientific expertise in thermo- and hydrodynamics, the behavior of gases under different conditions (and other areas of physics) in fact features practical implications that turned out to be extremely relevant for human powered flight. And the aforementioned, prominent mathematician-astronomer Simon Newcomb (1901) also argued in the early 1900s—in his article, "Is the airship coming?"—that the impossibility of flight was a scientific fact, as there was no combination of physical materials that could be combined to enable human flight (for historical details, see Anderson, 2004; Crouch, 2002).

The question then is: how does someone still—despite seemingly clear evidence and scientific consensus—hold onto a belief that appears delusional? In the case of human flight, the data, evidence, and scientific consensus were firmly against the possibility. No rational Bayesian should have believed in heavier-than-air flight. Furthermore, the evidence against it was not just empirical (in the form of LeConte's bird and other data) and based on science and scientific consensus (in the form of Kelvin and Newcomb's physics-related arguments), but it also was observationally salient. Many aviation pioneers not only failed and were injured, but some also died. For example, in 1896, the German aviation pioneer Otto Lilienthal died while attempting to fly, a fact that the Wright brothers were well acquainted with (as they subsequently studied his notebooks extensively). And in 1903—just nine weeks before the Wright brothers succeeded—the scientist Samuel Langley failed spectacularly in his attempts at flight, with large scientific and lay audiences witnessing the failures. Reporting on Langley's flight attempts, the *New York Times* (1903) estimated that it would take the "combined and continuous efforts of mathematicians and mechanicians from one million to ten million years" to achieve human powered flight.

We have of course opportunistically selected a historical example of a particular data-belief asymmetry where a seemingly delusional belief—one that went against existing data, evidence, and scientific consensus—turned out to be correct. Cognitive psychologists often engage in the "opposite" exercise where they point to situations where humans doggedly persist in holding delusional beliefs—despite clear evidence against those beliefs—due to biased information processing, selective perception or biased sampling of data (Festinger et al., 1956; Kahneman, 2011; Kunda, 1990; Pinker, 2021; though see Anglin, 2019). Clearly these types of biases exist and are problematic. However, again, we think that the other side of heterogeneous beliefs—beliefs that *presently* might appear delusional, but turn out to be correct—also needs to be explored. Our example of flight offers an instance of a far more generalizable process where data-belief asymmetries in fact are essential. Heterogenous beliefs and data-belief asymmetries are the lifeblood of new ideas, new forms of experimentation, and new knowledge—as we discuss next.

THEORY-BASED CAUSAL LOGIC AND COGNITION

Building on the aforementioned data-belief asymmetry, next we discuss the cognitive and practical process by which humans engage in forward-looking theorizing and causal reasoning that enables them to, in essence, go "beyond the data"—or more specifically, to go beyond existing data, to experiment and produce new data and novelty. We specifically highlight how this form of cognitive and practical activity differs from computational, data-driven, and information processing-oriented forms of cognition—the hallmarks of AI and computational AI we discussed above—and allows humans to "intervene" in the world in forward-looking fashion. Approaches that focus on data-driven prediction take and analyze the world *as it is* without recognizing the human capacity to experiment and to understand *why* (cf. Pearl and Mackenzie, 2018)—and to realize beliefs that presently seem implausible due to the (present) lack of data and evidence. We extend the example of heavier-than-air flight to offer a practical example of this point, in an effort to provide a unique window into a more generalized and ubiquitous process of what we call "theory-based causal logic."

Our foundational starting point—building on Felin and Zenger (2017)—is that cognitive activity is a form of theoretical or scientific activity.[21] That is, humans generate forward-looking theories that guide their perception, search, and action. As noted by Peirce, the human "mind has a natural adaptation to imagining correct theories of some kinds (…). If man had not the gift of a mind adapted to his requirements, he could not have acquired any knowledge" (1957: 71). As highlighted by our example of language, the meager linguistic inputs of a child can scarcely account for the vast outputs, thus pointing to a human generative capacity to theorize. The human capacity to theorize—to engage in novel problem solving and experimentation—has evolutionary origins and provides a highly plausible explanation for evolutionary leaps and the emergence of technology (Felin and Kauffman, 2023).

Importantly, theory-based cognition enables humans to *do* things. This is also the basis of the so-called "core knowledge" argument in child development (e.g., Carey and Spelke, 1996; Spelke et al., 1992). Humans develop knowledge like scientists, through a process of conjecture, hypothesis, and experimentation. While computational approaches to cognition focus on the primacy of data and environmental inputs, the theory-based view of cognition focuses on the active role of humans in not just learning about their surroundings but also their active role in experimentation, the generation of new knowledge, and innovation (Felin and Zenger, 2017). Without this active, generative, and forward-looking component of theorizing, it is hard to imagine how knowledge would grow—whether we are talking about individual, collective, or scientific knowledge. This is nicely captured in the title of an article in developmental psychology: "If you want to get ahead, get a theory" (Karmiloff-Smith and Inhelder, 1974). This also echoes Kurt Lewin's maxim, "there is nothing as practical as a good theory" (1943: 118). The central point here is that theories are not just for scientists. Theories are pragmatically useful for anyone seeking to understand and influence their

[21] A central aspect of this argument—which we unfortunately do not have room to explicate in this paper—is that humans are *biological* organisms. The theory-based view builds on the idea that all organisms engage in a form of forward-looking problem solving. A central aspect of this approach is captured by the biologist Rupert Riedl who argued that "Every conscious cognitive process will show itself to be steeped in theories; full of hypotheses" (1984: 8). To see the implications of this biological argument on human cognition—particularly in comparison to statistical and computational approaches—see Felin and Koenderink (2022; also see Roli et al., 2022; Jaeger et al., 2023). For the embodied aspects of human cognition, see Mastrogiorgio et al., 2022.

surroundings—*theories help us do things*. Theorizing is a central aspect of human cognitive and practical activity. Thus, as argued by Dewey, "the entities of science are not only from the scientist" and "individuals in every branch of human endeavor should be experimentalists" (1916: 438-442). We build on this intuition and extend it into new and novel domains, along with contrasting it with AI-informed models of cognition.

The theory-based view—in the context of decision making and strategy—extends the above logic and emphasizes the importance of theorizing and theories in economic contexts, with widespread implications for cognition (Felin and Zenger, 2017). The central idea behind the theory-based view is that economic actors can (and need to) develop unique, firm-specific theories. Theories do not attempt to map existing realities, but rather to generate unseen future possibilities. In economics, a roughly similar idea has been captured under the idea of "reverse Bayesianism" (see Karni and Vierø, 2013). Theories can be seen as a mechanism for "hacking" competitive factor markets (cf. Barney, 1986), enabling economic actors to see and search the world differently. Awareness for new possibilities is cognitively developed top-down (Felin and Koenderink, 2022). Theories also have central implications for how to efficiently organize or govern the process of realizing something that is new (Wuebker et al., 2023). This approach has been empirically tested and validated (e.g., Agarwal et al., 2023; Camuffo et al., 2021; Novelli and Spina, 2022), including important theoretical extensions (e.g., Ehrig and Schmidt, 2022; Zellweger and Zenger, 2022).[22] The practical implications of the theory-based view have also led to the development of managerial tools to assist startups, economic actors, and organizations in creating economic value (Felin, Gambardella, and Zenger, 2021).

Our goal in this section of the paper is *not* to exhaustively review the theory-based view. Rather, our goal now is to further build out the *cognitive* aspects of the theory-based view, and to contrast these with backward-oriented, data-focused approaches to cognition and AI. We highlight how the human capacity for theorizing and causal reasoning differs from AI's emphasis on data-driven prediction. A theory-based view of cognition allows humans to intervene in the world, beyond the given data—not just to process, represent, or extrapolate from existing data. Theories enable the generation of nonobvious data and new knowledge through experimentation. We highlight how our approach to cognition differs significantly from the arguments and prescriptions suggested by computational, Bayesian, and AI-inspired approaches to cognition.

[22] There are parallel literatures in strategy that focus on mental representations (e.g., Csaszar and Levinthal, 2016) and forward-looking search and representation (e.g., Gavetti and Levinthal, 2000; also see Gans, Stern and Wu, 2019).

It is important to carefully establish these differences, as AI-based and computational approaches—as extensively discussed at the outset of this paper—are said to replace human judgment and cognition (e.g., Kahneman, 2018).

Data-Belief Asymmetry Revisited

Heterogeneous beliefs provide the initiating impetus for the theory-based view of cognition. From our perspective, for beliefs to be a relevant concept for understanding cognition and decision making, beliefs do not necessarily—in the first instance—need to be based on data. We are specifically interested in forward-looking beliefs, beliefs that *presently* lack evidence or *even go against existing data*. Forward-looking beliefs, then, are more in search of data rather than based on data. At the forefront of knowledge, data is an outcome of beliefs—coupled with action and experimentation—rather than beliefs being an outcome of existing data.

The problem is that it is hard to ex ante distinguish between beliefs that indeed are delusional versus those that simply are ahead of their time. Data-belief asymmetry is critical in situations where data "lags" belief, that is, situations where the corroborating data simply has not yet been identified, found, or experimentally generated. In many cases, beliefs do not automatically verify themselves. Rather, more often than not they require some form of action and experimentation. The search for data in support of an uncommon, contrarian or discrepant belief necessarily looks like irrational motivated reasoning or confirmation bias (Kunda, 1990; cf. Hahn and Harris, 2014). To briefly illustrate, Galileo's belief in heliocentrism went against the established scientific data and consensus, and even plain common sense. Geocentric conceptions of the universe were observationally well-established. And they were successful: they enabled precise predictions about the movement of planets and stars. Even everyday observation verified that the earth does not move and that the sun seemingly circles the earth. Galileo's detractors essentially argued that Galileo was engaged in a form of biased, motivated reasoning against the Catholic Church, by trying to take humankind and the immovable Earth away from the center of God's creation.

Before discussing the actions associated with the realization of seemingly contrarian or delusional beliefs, it is worth emphasizing the important role of beliefs as motivators of action. Namely, the strength or degree of one's belief can be measured by one's likelihood to take action as a result of that belief (Ramsey, 1931; also see Felin, Gambardella, and Zenger, 2021). By way of contrast, the degree or strength of belief, based on probabilistic or Bayesian models of cognition (cf. Pinker, 2021), is tied to *existing data* and the *weight* of the *available* evidence (cf. Keynes, 1921), rather than the likelihood of taking action—a significant difference.

Notice the implications of this in a context of our earlier example, human powered flight. Beliefs played a central role in motivating action on the part of aviation pioneers *despite* overwhelming data and evidence against those beliefs. In a sense, those pursuing flight did not appropriately update their beliefs. Much if not most of the evidence was against them, but somehow they still believed in the plausibility. One of the Wright brothers, Wilbur, wrote to the scientist and aviation pioneer Samuel Langley in 1899 and admitted that "for some years I have been *afflicted with the belief* that flight is possible. My *disease* has increased in severity and I feel that it will soon cost me an increased amount of money if not my life" (Wright and Wright, 1881-1940, *emphasis added*). Wilbur clearly recognized that his belief about flight appeared delusional to others, as is evident from his many letters. But this belief motivated him to experiment and problem solve and to make the seemingly delusional belief a reality (only four short years later). Contrast the Wright brothers' belief with the belief of Lord Kelvin, one of the greatest scientific minds of the time. When invited to join the newly-formed Aeronautical Society a decade earlier, Kelvin declined and said "I have not the slightest molecule of faith in aerial navigation." Here Kelvin might have been channeling a scientific contemporary of his—the mathematician William Clifford—who argued that "it is wrong always, everywhere, and for anyone to believe anything on insufficient evidence" (2010: 79). Kelvin did not want to lend support to what he considered an anti-scientific endeavor. Without the slightest belief in the possibility of human flight, he naturally did not want to support anything that suggested human powered flight might be possible. But for the Wright brothers, the possibility of powered flight was very much a "live hypothesis" (James, 1967). *Despite* the data, they believed human flight might be possible, and took specific steps to realize their belief.

This approach presents problems for the very idea of rationality (cf. Chater et al., 2018; Felin and Koenderink, 2022). After all, to be a rational human being, our knowledge should be based on evidence. Knowledge should be proportionate to the evidence at hand. In a strict sense, the very concept of beliefs is not even needed, as one can instead simply talk about knowledge—that is, beliefs justified by or proportionate to the evidence at hand. This is succinctly captured by Pinker who argues "*I don't believe in anything you have to believe in*" (2021: 244). This seems like a reasonable stance. It is also the basis of Bayesian approaches where new data (somehow) emerges, and where we can update our beliefs and knowledge accordingly—providing us an "optimal way to update beliefs given new evidence" (Pilgrim et al., 2024). This is indeed the implicit stance of cognitive approaches that focus on computational and probabilistic belief updating (e.g., Dasgupta et al., 2020).

But data-belief asymmetries—where existing data presently does not corroborate beliefs, or even goes against them—can be highly useful, even essential. They are the raw materials of technological and scientific progress. Data-belief asymmetries direct our awareness toward new data and possible experiments to generate the evidence to support a belief. Of course, the idea of "seeking-data-to-verify-a-particular-belief" is the very definition of delusion and a host of associated biases, including confirmation bias, motivated reasoning, cherry picking, denialism, and belief perseverance.[23] To an outsider, this looks like the perfect example of "the bad habit of seeking evidence that ratifies a belief and being incurious about evidence that might falsify it" (Pinker, 2021: 13; also see Hahn and Harris, 2014). Belief in human powered flight readily illustrates this, as there was plenty of evidence to falsify the Wright brothers' belief in the plausibility of heavier-than-air flight. Holding an asymmetric belief seems to amount to "wishful thinking", or "protecting one's beliefs when confronted with new evidence" (Kruglanski, Jasko and Friston, 2020: 413; though see Anglin, 2019). The Wright brothers were continuously confronted with evidence that disconfirmed their belief, including Samuel Langley's public failures with flight or the knowledge of Lilienthal's failed attempts (and his death due to a failed flight attempt). But in these instances, ignoring the salient data and evidence—*not* updating beliefs based on seemingly strong evidence and even scientific consensus—turned out to be the correct course of action.

Now, perhaps the examples we have discussed represent rare and exceptional instances—instances that we can only discuss ex post, once the delusional or veridical nature of the beliefs in question is actually known. But again, scientists largely point to negative side of data-belief asymmetry all the time, by emphasizing situations where evidence is irrationally ignored by humans (e.g., Bordalo et al., 2024; Chater, 2018; Kahneman, 2011; Pinker, 2021; Scheffer et al., 2022). We are emphasizing the other side of the equation, where the present data is (rightly) ignored, and where this leads to positive outcomes. Belief-data asymmetries

[23] In economics there has similarly been an emphasis on how beliefs lead to negative outcomes. For example, Gennaioli and Shleifer's (2018) "theory of beliefs" focuses on beliefs that turn out to be delusional and are the result of poor judgment, biased information processing, and selective perception. In a related vein, Bordalo et al (2023) largely argue that humans are poor statisticians—selectively attending to and inappropriately weighting evidence and feedback—leading to suboptimal outcomes. Here we highlight discrepant beliefs that appear delusional and highly biased—to some, or even a majority, of actors—*in the present*, but turn out to be correct. Importantly, our theory is one of belief asymmetry rather than bounded rationality, bias, or information asymmetry (cf. Felin, Gambardella, Novelli and Zenger, 2024).

are important and they are central to human cognition. There are times when being (seemingly) irrational—ignoring evidence, disagreeing about its interpretation, or selectively looking for the right data—turns out to be the correct course of action. Human powered flight of course is a particularly vivid illustration of this, though even more mundane forms of human behavior are fundamentally characterized by a similar process (Felin and Koenderink, 2022). Most important for present purposes, our argument is that beliefs have a causal role of their own and can be measured by our propensity to act on them (Felin et al., 2020; Ramsey, 1931). Of course, having beliefs or having a willingness-to-act on them does not assure us that they are true. But they are an important motivation for action (Ajzen, 1991; Bratman, 1987).[24] And again, notice that our emphasis on beliefs should not be seen as an attempt to dismiss the importance of data. Rather, as we highlight next, we emphasize the role of beliefs and theory in directing awareness toward the relevant data or the relevant experiments for the generation of evidence.

From Beliefs to Problem Solving and Experimentation

The realization of beliefs is not automatic. A central aspect of beliefs is their propensity to lead to directed experimentation and problem solving. Beliefs enable actors to experiment, to generate or find novel data and solutions, which presently are hidden (Felin, Gambardella and Zenger, 2021). Put differently, if the current conditions of the world do not provide evidence to support the plausibility of a belief, agents can engage in directed experimentation and problem solving to realize their beliefs.

Our view of cognition and action here is more generally informed by the idea that theorizing can guide humans to develop an underlying *causal logic* that enables us to intervene in the world (Pearl and Mackenzie, 2018). This intervention-orientation means that we do not simply take the world as it is, rather we counterfactually think about possibilities and future states, with an eye toward taking specific action, experimenting and generating the *right* evidence. This shifts the locus from backward-oriented information processing and prediction (where the data is given), to doing and experimentation (where the right data and evidence is identified or generated). This involves actively questioning and manipulating causal structures, allowing for a deeper exploration of "what if" scenarios. Counterfactual thinking empowers humans to probe

[24] Beyond the work of Ramsey, Ajzen, and Bratman mentioned above, there is of course a large literature on how beliefs motivate action. Our emphasis here is on the interaction between data and beliefs (and eventually, the role of theory-based causal logic), as this has manifest in computational, Bayesian and probabilistic forms of AI and cognition.

hypothetical alternatives and dissect causal mechanisms offering insights into necessary and sufficient conditions for an outcome (Felin, Gambardella, Novelli, and Zenger, 2024). This approach is significantly different from input-output and information processing-oriented models of AI and computational cognition, and various "data-driven" or Bayesian approaches to decision making. AI-based models of cognition largely focus on *patterns* based on past associations and correlations—prediction is based on past data. But these approaches lack an ability to understand underlying causal structures, hypothetical possibilities, and possible interventions (cf. Felin et al., 2020; Pearl and Mackenzie, 2018). This is the role of theory-based causal logic.

A doing-orientation can be illustrated by extending our example of human powered flight. This example also aptly illustrates the difference between how data-oriented and evidence-based scientists thought about the possibility of human powered flight versus how more intervention-oriented and causal logic-based practitioners like the Wright brothers thought about it. To understand flight, the Wright brothers delved into the minutiae of *why* previous attempts at flight had not succeeded. While failed flight attempts and the death of Lilienthal (and others) were used by many as data to claim that flight was impossible, the Wright brothers looked at the specific *reasons* why these attempts had failed.[25] And while scientists had used bird data to argue that human flight was impossible (due to weight) (e.g., LeConte, 1888; Newcomb, 1901), the Wright brothers paid attention to a *different* aspect of birds flight. Ironically, bird-related data—again, different aspects of it—provided seeming evidence for both those advocating for and against flight. LeConte focused on the weight of birds, while the Wright brothers engaged in observational studies of the *mechanics* of bird flight and anatomy (why birds were able to fly), for example, carefully studying the positioning of bird wings when banking and turning.

The key difference was that the Wright brothers—with their belief in the plausibility of flight—were building a *causal theory of flying* rather than looking for data that confirmed or disconfirmed whether flight was possible. The Wright brothers ignored the data and the scientific arguments of the naysayers. From the Smithsonian, the Wright brothers requested and received details about numerous historical flight attempts, including Otto Lilienthal's records. The Wright brothers notes and letters reveal that they carefully studied the

[25] The Wright brothers respected Otto Lilienthal and carefully analyzed his data. Based on their own experimentation, they found that some of his data on "lift" overestimated lift coefficients. Lilienthal tested one wing shape while the Wright brothers experimented with various options. The Wright brothers constructed their own wind tunnel to gather aerodynamic data. Their tests led them to develop new lift, drag, and pressure distribution data, which differed from Lilienthal's findings. This data was critical in designing their successful aircraft.

flight attempts and aircraft of earlier pioneers like George Cayley, Alphonse Penaud, and Octave Chanute (Anderson, 2002; McCoullough, 2015; Wright and Wright, 1841-1940). They studied various aspects of past flight attempts: the types of airplanes used, details about wing shape and size, weather conditions, and underlying aerodynamic assumptions.

All of this study led to the development of their own theory of flying. Their theory consisted not just of a contrarian belief, but the gradual specification of an underlying causal logic of flight, which included the articulation of the specific problems they needed to solve for human powered flight to be possible. The Wright brothers reasoned that it was essential to solve three problems related to flight—namely (a) lift, (b) propulsion, and (c) steering. To illustrate the power of developing a theory-based causal logic, and identifying specific problems to solve, coupled with directed experimentation, we briefly discuss how they addressed one of the problems, the problem of lift.

In terms of lift, the Wright brothers understood that to achieve flight they needed a wing design that could provide sufficient lift to overcome the weight of their aircraft. Indeed, prominent scientists had argued that the prohibiting factor of human flight was weight (again, pointing to insect flight and the weight of those birds that fly and those that do not). The Wright brothers felt that the concern with weight was not insurmountable. Informed by their investigations into bird flight (and the flight attempts of others), they approached this problem through a series of experiments that included the construction and testing of various airfoils. Their experimentation was highly targeted and data-oriented, testing various wing shapes, sizes, and angles. They also quickly realized that not everything needed to be tested at scale and that their experiments with lift could more safely and cost effectively be done in laboratory conditions. Thus they constructed their own wind tunnels. Targeted tests within these tunnels allowed the Wright brothers to learn the central principles of lift. They measured everything and kept meticulous track of their data—data that they generated through ongoing experimental manipulation and variation. This hands-on experimentation allowed them to collect data on how different shapes and angles of attack affected lift. By systematically varying these parameters and observing the outcomes, they were effectively employing causal reasoning to identify the conditions under which lift could be maximized. Their discovery and refinement of wing warping for roll control was a direct outcome of understanding the causal relationship between wing shape, air pressure, and lift.

The same processes of causal logic, experimentation, and problem solving were also central for solving propulsion and steering or control. And more generally, the Wright brothers were careful scientists in every aspect of their attempt to realize their belief in human powered flight. For example, to determine a suitable place for their flight attempts, they contacted the US Weather Bureau. They had established what the optimal conditions might be for testing flight. They needed four things: consistent wind (direction and strength), wide open spaces, soft or sandy landing surfaces, and privacy. They received several suggestions from the US Weather Bureau and chose Kitty Hawk, North Carolina for the site of their "real world" trial (Wright and Wright, 1881-1940).

The Wright brothers' approach to flight offers a useful case study and microcosm of how theory-based causal logic enables belief-realization, even when beliefs seemingly are not supported by existing data, evidence, or science. Based on their extensive study and theorizing, the Wright brothers engaged in directed experimentation and problem solving—to solve the central problems of lift, propulsion and steering. Their approach exemplifies the application of causal logic to understand and manipulate the world. Their success was not just in achieving flight but in demonstrating how a systematic, intervention-based approach can unravel the causal mechanisms underlying complex phenomena and overcome the shortcomings of existing data.

As is implied by our arguments, we think the economic domain is replete with opportunities for those with asymmetric beliefs to develop theory-based causal logics and engage in directed experimentation and problem solving (Felin and Zenger, 2017). As we have argued, existing theories of cognition are overly focused on data-belief symmetry rather than data-belief asymmetry and how the latter enables the emergence of heterogeneity and the creation of novelty and value. Data-belief symmetry is inherent to AI-based models that focus on prediction based on past data. Next we further explore the implications of this argument for decision making under uncertainty and strategy.

DISCUSSION: THE LIMITS OF PREDICTION FOR DECISION MAKING UNDER UNCERTAINTY

As we have extensively discussed in this article, AI and the cognitive sciences use many of the same metaphors, tools, methods, and ways of reasoning about intelligence, rationality, and the mind. The prevailing assumption in much of the cognitive sciences is that the human mind is a computational, input-output system

(Christian and Griffiths, 2016). Computational and algorithmic systems emphasize the power of *prediction* based on past data. The centrality of prediction is echoed by one the pioneers of AI, Yann LeCun (2017), who argues that "prediction is the essence of intelligence."

Clearly the predictive capabilities of AI are powerful. But is prediction the central for decision making *under uncertainty* as well (that is, in unpredictable situations)? Many argue that this is the case (e.g., Davenport and Kirby, 2017, Kahneman, 2018). For example, in their widely-acclaimed book *Prediction Machines*: *The Simple Economics of Artificial Intelligence*—Agrawal, Gans, and Goldfarb emphasize that, stripped down to its essence, "AI is a *prediction* technology" (2022: 22-32). And the central claim of their book is that "prediction is at the heart of making decisions *under uncertainty*." (2022: 7, *emphasis added*). One way to summarize our argument in this paper is that we disagree with the importance placed on prediction—particularly in the form it is manifest in AI—especially in situations of uncertainty. Since the emphasis on prediction is commonplace, it is worth carefully pinpointing *why* we disagree with the importance placed on prediction.

When it comes to prediction, Agrawal et al's argument might be summarized by pointing to a relatively common causal chain (of sorts), one that proceeds from data to information to prediction and to a decision, or in short: data→ information→prediction→decision.[26] They argue that "data provides the information that enables a prediction" and prediction in turn is "a key input into our decision making." This causal chain—from data to information to prediction and decision—certainly has intuitive appeal and mirrors what AI systems are good at: taking in vast amounts of inputs and data, processing this information, and then making predictions that can be used to make decisions. In short, as emphasized by Agrawal et al (2022) and many others, data-driven prediction is at the heart of not just language models but AI more generally, and also placed center stage in cognition.

But as we have highlighted throughout this paper, the problem is that data—data that is *presently* available or given—is not likely to be the best source of information and prediction when making forward-looking decisions. Data is snapshot of the past. Even vast amounts of data are unlikely to somehow enable one to anticipate the future (see Felin, Kauffman, Koppl and Longo, 2014). What is needed is some mechanism for projecting into the future and identifying the *relevant* data and evidence, or experimentally

[26] This has parallels with the data-information-knowledge-wisdom or "DIKW" framework. For discussions of this see Felin, Koenderink, Krueger, Noble and Ellis, 2021, and Yanai and Lercher, 2020.

generating new data. This is the role of a theory, which is a critical element that is missing from data-first and prediction-oriented approaches to AI and cognition. We grant that for various routine and repetitive decisions, prediction undoubtedly is a useful tool. Data-based prediction can be highly powerful in predictable situations, situations that match or extrapolate from the past. This matches what AI and prediction-based cognition is really good at, namely, the *minimization* of surprise and reduction of error. More broadly this also matches the strong emphasis that many scholars of judgment and decision making put on "consistency"—and the eagerness to avoid noise (Kahneman, Sibony and Sunstein, 2021).

But many types of decision making are not about uncertainty reduction through error minimization using existing data. The purpose of large swaths of decision making is more about (in a sense) *maximizing* surprise and error, or what to others might look like error. In a strategy context, the most impactful opportunities and sources of value are not founded on immediately-available data. Rather, important decisions like this require the development of a theory, founded on some kind of heterogeneous belief, that maps a causal path or logic—for how to test the theory, experiment, and gather new evidence—to realize the belief. In an important sense, strategic decision making has more to do with unpredictability and the maximization of surprise rather than prediction and the minimization of surprise. Some decisions are highly-impactful, low-frequency and rare, and fraught with uncertainty (Camuffo et al., 2022), and therefore not amenable to algorithmic processing. This is why theory-based decision making is not about appropriately representing the structure of the environment, or about bounded rationality or listening to customers—rather it is about developing a forward-looking theory and causal logic about how to experiment and create value (Felin, Gambardella, Novelli and Zenger, 2024).

Notice that our focus on unpredictability and surprise does *not* mean that we are somehow outside the realms of science or data. Quite the contrary. *The process of making forward-looking decisions is about developing an underlying theory-based causal logic of how one might intervene in the world to create salience for new data through experimentation and problem solving.* As put by Einstein, "whether you can observe a thing or not depends on the theory which you use. It is the theory which decides what can be observed" (Polanyi, 1974: 604). Salience to the right (or new) data and experiment is given by a theory, not by past data. In this sense, theories could be said to have a "predictive" function, though here prediction is not a data-driven or error-minimizing process as it has been defined and operationalized within the context of AI (Agrawal et al., 2022) and cognitive science (cf. Clark, 2018). Now, if the task at hand is mundane—for example, "predict the next word in this sentence" or "tell me what you expect to see next"—then prediction with existing data can be useful. But the theory-based perspective of cognition is more focus on the forward-looking aspects of cognition, and how human agents realize beliefs by developing a multi-step causal path that enables the realization of beliefs through problem solving and experimentation. This is precisely what our example of the Wright brothers' theory of flying—and their experimentation and problem solving—illustrates. It serves as microcosm of a far more general process of how humans cognitively intervene in their surroundings and through experimentation realize their beliefs. The economic domain is full of examples of how economic actors engage in this process (Felin and Zenger, 2017).

Our emphasis on surprise and unpredictability—rather than predictability and the minimization of error—is particularly important in competitive contexts. If everyone has access to the same prediction machines and AI-related information processing tools, then the outcomes are likely to be homogeneous. Strategy—if it is to create value—needs to be unique and firm-specific. And this firm-specificity is tied to unique beliefs and the development of a theory-based logic for creating value that is unforeseen (not predictable) by others. This assists economic actors in their attempts to "hack" competitive factor markets (Barney, 1986), to develop unique expectations about the value of assets and activities. It also enables firms to "search" in a more targeted fashion, rather than engaging in costly and exhaustive forms of global search (Felin, Kauffman and Zenger, 2023). Prediction based engines, while there are attempts to fine-tune them, are inherently based on past frequencies, correlations, and averages, rather than extremes. And in many instances, it is the extremes that turn out to be far more interesting, as these provide the seeds of the (eventual) beliefs and data that we take for granted.

In all, we disagree with the emphasis that has been placed on prediction, algorithmic processing, and computation in decision making and cognition (e.g., Agrawal et al., 2022; Christian and Griffiths, 2016). Human decision making should not be relegated to AI (Kahneman, 2018). AI and AI-inspired models of cognition are based on backward-looking data and prediction rather than any form of forward-looking theory-based causal logic. Emphasizing or relying on data and prediction is a debilitating limitation for not just decision making and cognition, but also for understanding knowledge generation and even scientific progress. Therefore we have emphasized the importance of heterogenous beliefs in human cognition, and the development of theory-based causal logic that enables problem solving, experimentation and the generation of new data.

FUTURE RESEARCH OPPORTUNITIES

The above arguments suggest a number of research opportunities, particularly when it comes to understanding strategy and decision making under uncertainty. First, there is an opportunity to study *when* and *how* AI-related tools might be utilized by humans (like economic actors) to create new value or to aid in decision making. If AI—as a cognitive tool—is to be a source of competitive advantage, it has to be utilized in *unique or firm-specific* ways. AI that uses universally-available training data will necessarily yield generic and non-specific outputs. There is the risk that off-the-shelf AI solutions will be susceptible to the "productivity paradox" of information technology (Brynjolffson and Hitt, 1998), where investments in AI actually do *not* yield any gains to those buying these tools (rather, only to those selling these technologies). Thus there is an opportunity to study how a specific decision maker's—like a firm's—own theory of value can drive the process of AI development and adoption. For AI to actually be a useful tool for strategy and decision making, AI needs to be customized, purpose-trained, and fine-tuned—it needs to be made *specific*—to the theories, datasets, and proprietary documents of decision makers like firms. For example, advances on "retrieval-augmented generation" seem to offer a promising avenue to enhance specificity when using AI in strategic decision-making. Any adoption of AI should be deliberate about which corpora and training data are utilized (and which not) when seeking unique AI-driven outputs. After all, the outputs of an AI—tailored to use specific data—are also the product of human agents who make decisions about which data are relevant and (which are not) for the decision at hand. It is here that we see an opportunity to understand how humans might *uniquely* interact with AI to generate these tools and associated human-AI interfaces. Early work has begun to look at how firms utilize AI to increase innovation or how various human-AI hybrid solutions enable better decision making (e.g., Babina et al., 2024; Bell et al., 2024; Choudhary, Marchetti, Shrestha, and Puranam, 2023; Girotra et al., 2023; Gregory et al., 2021; Kemp, 2023; Kim et al., 2023; Raisch and Fomina, 2023). But there are promising opportunities to study how a particular economic actor's or firm's own *theory* and causal logic—as well as their unique or firm-specific sources of data and information—can shape the development or adoption of AI-related tools for executing strategy and making decisions.

Second, there are ongoing opportunities to research—and develop taxonomies of—the respective capabilities of humans versus AI when it comes to *different types* of tasks, problems, and decisions. There is much excitement, hype, and fearmongering about the prospects of AI replacing humans *tout court* (cf. Grace et al., 2024). However, in reality, there will likely be a division of labor between humans and AI—with each focusing on the types of tasks, problems, and decisions that it is best suited for. There is an opportunity to study how economic actors like firms contingently "match" humans (and their cognitive capacities, jobs, roles) versus algorithms (or AI-related tools) with the right tasks and decisions. At present, clearly AI is remarkably well-suited for tasks and decisions that are repetitive, computationally intensive, and that directly extrapolate from past data. A significant number of decisions made by humans are relatively routine and amenable to algorithmic processing (Amaya and Holweg, 2024; Holmström, Holweg, Lawson, Pil, and Wagner, 2019). AI will therefore undoubtedly play a key role in many areas of management, especially where processes repeat, such as operations (Amaya and Holweg, 2024; Holmström, Holweg, Lawson, Pil, and Wagner, 2019). However, some decisions are more low-frequency and rare (Camuffo et al., 2022), and therefore not amenable to AI. Here we anticipate that humans will continue to play a central role, given their ability to engage in forward-looking theorizing and the development of causal logic beyond extant data. That said, naturally there is a "sliding scale" (and interfaces) between routine and non-routine decision making. Even in the context of rare-and-highly-impactful decision making, AI might play a role, perhaps in augmenting humans in the gathering, processing, or aggregation of information. As we have discussed in this paper, AI and humans have their respective strengths and limitations. Existing work tends to compare AI and humans on the same benchmarks, rather than recognizing the respective strengths of each. Studying the comparative capabilities of AI and humans—their respective capabilities, limitations, and ongoing evolution—represents a significant opportunity for future work.

Third, our arguments point to perhaps more "foundational" questions about the very nature of humans, particularly related to the purportedly computational nature of human cognition. While questions about the nature of cognition might sound overly abstract and philosophical, they are critically important as they have downstream consequences for the assumptions we make, the methods we employ, as well as how and what we study. Here we echo Herbert Simon who argued that "*nothing* is more important in setting our research agenda and informing our research methods than *our view of the nature of the human beings* whose behavior we are studying" (1985: 303; *emphasis added*). So, what is the predominant view of human cognition within AI and the cognitive sciences (and by extension, in economics and strategy)? The predominant view of humans is that they are input-output devices engaged in information processing, akin to computers. The computer has served as a central, organizing metaphor of human cognition for well over seven decades—from the work of Alan Turing and Herbert Simon to modern instantiations of artificial neural networks, predictive processing, and the Bayesian brain (e.g., Cosmides and Tooby, 2013; Knill and Pouget, 2004; Goldstein and Gigerenzer, 2024; Kotseruba and Totsos, 2020; Russell and Norvig, 2022; Sun, 2023). A generalized computational approach to cognition, however, does not take into consideration the comparative *nature* of the organism under study, because humans, organisms, and machines are seen as "invariant" (see Simon, 1990; cf. Gershman et al., 2015; Simon, 1980). Studying these differences represents a significant opportunity for future work. There are significant differences in cognition, and these differences deserve careful attention. For example, computers do not meaningfully make decisions about *which* inputs might be relevant and which are not (nor can they meaningfully identify a new input), while humans have control over which inputs they might select or "generate" in the first place (e.g., Brembs, 2021; Felin and Koenderink, 2022; Yin, 2020). Human cognition is a form of forward-looking theorizing which is oriented toward experimentation and problem solving. Notice that we are not trying to argue for some kind of human exceptionalism, as these capacities are manifest—in different ways—across biological organisms more broadly (Riedl, 1984; cf. Popper, 1991).[27] There are significant research opportunities to study the endogenous and comparative factors that enable biological organisms and economic agents to theorize, problem solve, experiment—and to compare various forms of biological intelligence with artificial and nonbiological forms (cf. Levin, 2024). Treating all intelligence as generalized computation unnecessarily narrows the scope of theoretical and empirical work, and fundamentally misses the rich and heterogeneous ways that intelligence manifests itself across systems. Furthermore, the interfaces between biological and nonbiological forms of intelligence—as is manifest in the human use of technology and tools in evolution (Felin and Kauffman, 2023)—provide intriguing opportunities for future work.

[27] For example, even simple organisms like *Drosophila* (fruit flies) exhibit novel and surprising behaviors—like initiating activity, expectations, and problem solving—that cannot be explained by or reduced to environmental inputs, genetic factors or neural processing (see Heisenberg, 2014). Recent experiments highlight how rodents develop and test hypotheses and use cognitive strategies (Zhu and Kuchibhotla, 2024).

CONCLUSION

In this article we have focused on the *differences* in cognition between AI and humans. While AI-inspired models of cognition continue to emphasize the similarities between machines and humans, we argue that AI's data and prediction-orientation does not capture human cognition. We grant that there are some parallels between AI and human cognition, as (very) broad forms of information processing. But we specifically emphasize the forward-looking nature of human cognition and how theory-based causal logic allows humans to intervene in the world, to engage in directed experimentation, and develop new knowledge and novelty. Heterogeneous beliefs and theories—data-belief asymmetries—enable the identification or generation of *new* data (for example, through experimentation), rather than merely being reliant on prediction based on the past data. AI-based computational models are necessarily built on data-belief symmetries. AI cannot causally map and project into or anticipate the future, as illustrated by LLMs which are delimited by past data. That said, our arguments by no means negate or question many of the exciting developments within the domain of AI. We anticipate that AI will help humans make better decisions across many domains, especially in settings that are conducive to routine, repetition, and computation. However, decisions under uncertainty—given the emphasis on unpredictability, surprise, and *the new*—provide a realm that is not readily amenable to data- or frequency-based prediction and associated computation. Thus we question the idea that AI will (or should) replace human decision making (e.g., Kahneman, 2018). We argue that humans—compared to computers and AI—have unique cognitive capacities that center on forward-looking beliefs and theorizing: the ability to engage in novel causal reasoning, problem solving, and experimentation.

REFERENCES

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., & McGrew, B. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agarwal, R., Bacco, F., Camuffo, A., Coali, A., Gambardella, A., Msangi, H., & Wormald, A. 2023. Does a theory-of-value add value? Evidence from a randomized control trial with Tanzanian entrepreneurs. *SSRN working paper*.

Aggarwal, C. C. 2018. *Neural networks and deep learning*. Springer Publishing.

Agrawal, A., Gans, J., & Goldfarb, A. 2022. *Prediction machines (updated and expanded): The simple economics of artificial intelligence*. Harvard Business Review Press.

Agrawal, A., Gans, J., & Goldfarb, A. 2022. *Power and prediction: The disruptive economics of artificial intelligence*. Harvard Business Review Press.

Agrawal, A., McHale, J., & Oettl, A. 2023. Superhuman science: How artificial intelligence may impact innovation. *Journal of Evolutionary Economics*, 33(5), 1473-1517.

Agrawal, A., McHale, J., & Oettl, A. 2024. Artificial intelligence and scientific discovery: A model of prioritized search. *Research Policy*, 53(5), 104989.

Ajzen, I. 1991. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179-211.

Amaya, J. & Holweg, M. 2024. Using algorithms to improve knowledge work. *Journal of Operations Management*, forthcoming.

Ananthaswamy, A. 2022. DeepMind AI topples experts at complex game Stratego. *Nature*.

Anderson, J. D. 2004. *Inventing Flight: The Wright brothers and their predecessors*. Johns Hopkins University Press.

Anderson, J. R. 1976. *Language, memory, and thought*. Psychology Press.

Anderson, J. R. 1990. *The adaptive character of thought*. Psychology Press.

Anglin, S. M. 2019. Do beliefs yield to evidence? Examining belief perseverance vs. change in response to congruent empirical findings. *Journal of Experimental Social Psychology*, 82, 176-199.

Babina, T., Fedyk, A., He, A., & Hodson, J. 2024. Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics*, 151, 103745.

Baker, B., Lansdell, B., & Kording, K. P. 2022. Three aspects of representation in neuroscience. *Trends in Cognitive Sciences*.

Bell, J. J., Pescher, C., Tellis, G. J., & Füller, J. 2024. Can AI help in ideation? A theory-based model for idea screening in crowdsourcing contests. *Marketing Science,* 43(1), 54-72.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Harari, Y. N., Hadfield, G., Russell, S., Kahneman, D., & Mindermann, S. 2023. Managing ai risks in an era of rapid progress. *arXiv preprint arXiv:2310.17688*.

Bengio, Y., Lecun, Y., & Hinton, G. 2021. Deep learning for AI. *Communications of the ACM*, *64*(7), 58-65.

Bhatia, S. 2023. Inductive reasoning in minds and machines. *Psychological Review*.

Biber, D. 1991. *Variation across speech and writing*. Cambridge University Press.

Binz, M., & Schulz, E. 2023. Turning large language models into cognitive models. *arXiv preprint arXiv:2306.03917*.

Binz, M., Dasgupta, I., Jagadish, A. K., Botvinick, M., Wang, J. X., & Schulz, E. 2023. Meta-learned models of cognition. *Behavioral and Brain Sciences*, 1-38.

Bratman M. 1987. *Intention, Plans and Practical Reason*. Harvard University Press: Cambridge, MA.

Bordalo, P., Conlon, J. J., Gennaioli, N., Kwon, S. Y., & Shleifer, A. 2023. How people use statistics (No. w31631). *National Bureau of Economic Research*.

Bory, P. 2019. Deep new: The shifting narratives of artificial intelligence from Deep Blue to AlphaGo. *Convergence*, 25(4), 627-642.

Brembs, B. 2021. The brain as a dynamically active organ. *Biochemical and Biophysical Research Communications*, *564*, 55-69.

Brynjolfsson, E., & Hitt, L. M. 1998. Beyond the productivity paradox. *Communications of the ACM*, 41(8), 49-55.

Buchanan, B. G., & Shortliffe, E. H. 1984. *Rule based expert systems: the mycin experiments of the stanford heuristic programming project*. Addison-Wesley Longman Publishing Co.

Buckner, C. J. 2023. *From deep learning to rational machines: What the history of philosophy can teach us about the future of artificial intelligence*. Oxford University Press.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., & VanRullen, R. 2023. Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

Camuffo, A., Cordova, A., Gambardella, A., & Spina, C. 2020. A scientific approach to entrepreneurial decision making: Evidence from a randomized control trial. *Management Science*, *66*(2), 564-586.

Camuffo, A., Gambardella, A., & Pignataro, A. 2022. Microfoundations of low-frequency high-impact decisions. *SSRN working paper.*

Carey, S., & Spelke, E. 1996. Science and core knowledge. *Philosophy of Science*, 63(4), 515-533.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., & Xie, X. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3), 1-45.

Chater, N. 2018. *Mind is flat: The remarkable shallowness of the improvising brain.* Yale University Press.

Chater, N., Felin, T., Funder, D. C., Gigerenzer, G., Koenderink, J. J., Krueger, J. I., & Todd, P. M. 2018. Mind, rationality, and cognition: An interdisciplinary debate. *Psychonomic Bulletin & Review*, *25*, 793-826.

Chater, N., Zhu, J. Q., Spicer, J., Sundh, J., León-Villagrá, P., & Sanborn, A. 2020. Probabilistic biases meet the Bayesian brain. *Current Directions in Psychological Science*, 29(5), 506-512.

Christian, B., & Griffiths, T. 2016. *Algorithms to live by: The computer science of human decisions.* Macmillan.

Chollet, F. 2019. *On the measure of intelligence.* arXiv preprint arXiv:1911.01547.

Chomsky N. 1975. *Reflections on language.* Pantheon: New York.

Choudhary, V., Marchetti, A., Shrestha, Y. R., & Puranam, P. 2023. Human-AI ensembles: When can they work? *Journal of Management.*

Clark, A. 2018. A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenology and the Cognitive Sciences*, *17*(3), 521-534.

Clifford, W. K. 2010. *The ethics of belief and other essays.* Prometheus Books.

Clough, D. R., & Wu, A. 2022. Artificial intelligence, data-driven learning, and the decentralized structure of platform ecosystems. *Academy of Management Review*, 47(1), 184-189.

Coady, C. A. J. 1992. *Testimony: A philosophical study.* Oxford University Press.

Constant, A., Friston, K. J., & Clark, A. 2024. Cultivating creativity: predictive brains and the enlightened room problem. *Philosophical Transactions of the Royal Society B*, 379(1895), 20220415.

Cosmides, L. and Tooby, J., 2013. Evolutionary psychology: New perspectives on cognition and motivation. *Annual Review of Psychology*, *64*, 201-229.

Crouch, T. D. 2002. *A dream of wings: Americans and the airplane, 1875-1905.* WW Norton & Company.

Csaszar, F.A., Ketkar, H.J. & Kim, H. 2024. AI and strategic decision making. Working paper presented at Bocconi University. October 2023.

Csaszar, F. A., & Levinthal, D. A. 2016. Mental representation and the discovery of new strategies. *Strategic Management Journal,* 37(10), 2031-2049.

Csaszar, F. A., & Steinberger, T. 2022. Organizations as artificial intelligences: The use of artificial intelligence analogies in organization theory. *Academy of Management Annals*, 16(1), 1-37.

Dasgupta, I., Schulz, E., Tenenbaum, J. B., & Gershman, S. J. 2020. A theory of learning to infer. *Psychological Review*, 127(3), 412.

Davenport, T. H., & Kirby, J. 2016. *Only humans need apply: Winners and losers in the age of smart machines*. New York: Harper Business.

Dewey, J. 1916. *Essays in experimental logic*. University of Chicago Press.

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., & Cui, C. 2022. Glam: Efficient scaling of language models with mixture-of-experts. *International Conference on Machine Learning* (5547-5569). PMLR.

Ehrig, T., & Schmidt, J. 2022. Theory-based learning and experimentation: How strategists can systematically generate knowledge at the edge between the known and the unknown. *Strategic Management Journal*, *43*(7), 1287-1318.

Ellis, K. 2024. Human-like few-shot learning via bayesian reasoning over natural language. *Advances in Neural Information Processing Systems*, *36*.

Feigenbaum, E. A. 1963. Artificial intelligence research. *IEEE Transactions in Information Theory*, *9*(4), 248-253.

Felin, T., Gambardella, A., Novelli, E., & Zenger, T. 2024. A scientific method for startups. *Journal of Management*.

Felin, T., Gambardella, A., & Zenger, T. 2021. Value lab: a tool for entrepreneurial strategy. *Management & Business Review*.

Felin, T., Kauffman, S., Koppl, R., & Longo, G. 2014. Economic opportunity and evolution: Beyond landscapes and bounded rationality. *Strategic Entrepreneurship Journal*, *8*(4), 269-282.

Felin, T., & Kauffman, S. 2023. Disruptive evolution: Harnessing functional excess, experimentation, and science as tool. *Industrial and Corporate Change*, 32(6): 1372-1392.

Felin, T., Kauffman, S., & Zenger, T. 2023. Resource origins and search. *Strategic Management Journal*, *44*(6), 1514-1533.

Felin, T., Koenderink, J., & Krueger, J. I. 2017. Rationality, perception, and the all-seeing eye. *Psychonomic Bulletin & Review*, *24*, 1040-1059.

Felin, T., Koenderink, J., Krueger, J. I., Noble, D., & Ellis, G. F. 2021. The data-hypothesis relationship. *Genome Biology*, 22(1), 1-6.

Felin, T., & Koenderink, J. 2022. A generative view of rationality and growing awareness. *Frontiers in Psychology*, *13*, 807261.

Felin, T., & Zenger, T. R. 2017. The theory-based view: Economic actors as theorists. *Strategy Science*, 2(4), 258-271.

Festinger, L., Riecken, H. W., & Schachter, S. 1956. *When prophecy fails*. University of Minnesota Press.

Franceschelli, G., & Musolesi, M. 2023. On the creativity of large language models. *arXiv preprint arXiv:2304.00008*.

Friedman, J. H., & Popescu, B. E. 2008. Predictive learning via rule ensembles. *Annals of Applied Statistics*, 2(3): 916-954

Friston, K., & Kiebel, S. 2009. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521), 1211-1221.

Friston, K. J., Da Costa, L., Tschantz, A., Kiefer, A., Salvatori, T., Neacsu, V., & Buckley, C. L. 2023. Supervised structure learning. *arXiv preprint arXiv:2311.10300*.

Friston, K. J., Ramstead, M. J., Kiefer, A. B., Tschantz, A., Buckley, C. L., Albarracin, M., & René, G. 2024. Designing ecosystems of intelligence from first principles. *Collective Intelligence*, *3*(1), 26339137231222481.

Gans, J. S., Stern, S., & Wu, J. 2019. Foundations of entrepreneurial strategy. *Strategic Management Journal*, *40*(5), 736-756.

Gavetti, G., & Levinthal, D. 2000. Looking forward and looking backward: Cognitive and experiential search. *Administrative Science Quarterly*, *45*(1), 113-137.

Gennaioli, N., & Shleifer, A. 2018. *A crisis of beliefs: Investor psychology and financial fragility*. Princeton University Press.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. 2015. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273-278.

Gigerenzer, G., & Goldstein, D. G. 2024. Herbert Simon on the birth of the mind-computer metaphor. In *Elgar Companion to Herbert Simon*. Edward Elgar Publishing.

Gingerich, O., & MacLachlan, J. 2005. *Nicolaus Copernicus: Making the earth a planet*. Oxford University Press.

Girotra, K., Meincke, L., Terwiesch, C., & Ulrich, K. T. 2023. Ideas are dimes a dozen: Large language models for idea generation in innovation. *Available at SSRN 4526071*.

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. 2024. Thousands of AI authors on the future of AI. *arXiv preprint arXiv:2401.02843*.

Gregory, R. W., Henfridsson, O., Kaganer, E., & Kyriakou, H. 2021. The role of artificial intelligence and data network effects for creating user value. *Academy of Management Review*, 46(3), 534-551.

Griffin, D., & Tversky, A. 1992. The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24(3), 411-435.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. 2010. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8), 357-364.

Gigerenzer, G. 2020. How to explain behavior? *Topics in Cognitive Science*, *12*(4), 1363-1381.

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D. & Paul, T. D. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, *26*(2), 248-265.

Goldman, A. I. 1999. *Knowledge in a social world.* Oxford University Press.

Goodfellow, I., Bengio, Y. & Courville, A., 2016. *Deep learning*. MIT Press.

Goyal, A., & Bengio, Y. 2022. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266), 20210068.

Hahn, U., & Harris, A. J. 2014. What does it mean to be biased: Motivated reasoning and rationality. In *Psychology of learning and motivation* (Vol. 61, pp. 41-102). Academic Press.

Hahn, U., Merdes, C., & von Sydow, M. 2018. How good is your evidence and how would you know? *Topics in Cognitive Science*, 10(4), 660-678.

Halliday, M.A.K. 1989. *Spoken and written language.* Oxford University Press.

Harari, Y. N. 2018. *21 Lessons for the 21st Century*. Random House.

Hart, B., and Risley, T. R. 2003. The early catastrophe: The 30 million word gap by age 3. *American Educator*, *27*(1), 4-9.

Hasson, U., Nastase, S. A., & Goldstein, A. 2020. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3), 416-434.

Hebb, D. O. 1949. *The organization of behavior: A neuropsychological theory*. Psychology Press.

Heisenberg, M. 2014. The beauty of the network in the brain and the origin of the mind in the control of behavior. *Journal of Neurogenetics,* 28(3-4), 389-399.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., & Couzin, I. D. 2015. Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46-54.

Hinton, G. E. 1992. How neural networks learn from experience. *Scientific American*, *267*(3), 144-151.

Hinton, G.E. 2023. Will digital intelligence replace biological intelligence? *University of Toronto Lecture.* [Available online here.]

Hinton, G. E., Osindero, S., & Teh, Y. W. 2006. A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527-1554.

Hohwy, J. 2013. *The predictive mind.* OUP Oxford.

Hohwy, J. 2020. New directions in predictive processing. *Mind & Language*, 35(2), 209-223.

Holmström, J., Holweg, M., Lawson, B., Pil, F.K. and Wagner, S.M., 2019. The digitalization of operations and supply chain management: Theoretical and methodological implications. *Journal of Operations Management*, 65(8), pp.728-734.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Hong, P., Ghosal, D., Majumder, N., Aditya, S., Mihalcea, R., & Poria, S. 2024. Stuck in the quicksand of numeracy, Far from AGI Summit: Evaluating LLMs' mathematical competency through ontology-guided pserturbations. *arXiv preprint arXiv*:2401.09395.

Jaeger, J., Riedl, A., Djedovic, A., Vervaeke, J., & Walsh, D. 2023. Naturalizing Relevance Realization: Why agency and cognition are fundamentally not computational. Working paper.

James, W. 1967. The writings of William James: A comprehensive edition. University of Chicago Press.

Johnson-Laird, P. N. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., & Hassabis, D. 2021. Highly accurate protein structure prediction with AlphaFold. nature, 596(7873), 583-589.

Kahneman, D. 2003. Maps of bounded rationality: Psychology for behavioral economics. *American Eonomic Review*, *93*(5), 1449-1475.

Kahneman, D. 2011. *Thinking fast and slow. Farrar*, Straus & Giroux.

Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. 2016. Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review*, 94(10), 38-46.

Kahneman, D. 2018. A Comment on Artificial Intelligence and Behavioral Economics. In *The Economics of Artificial Intelligence: An Agenda* (pp. 608-610). University of Chicago Press.

Kahneman, D., Sibony, O., & Sunstein, C. R. 2021. *Noise: A flaw in human judgment*. Hachette Publishing.

Karmiloff-Smith, A., & Inhelder, B. 1974. If you want to get ahead, get a theory. *Cognition*, *3*(3), 195-212.

Karni, E., & Vierø, M. L. 2013. Reverse Bayesianism: A choice-based theory of growing awareness. *American Economic Review*, *103*(7), 2790-2810.

Kemp, A. 2023. Competitive advantages through artificial intelligence: Toward a theory of situated AI. *Academy of Management Review*.

Keynes, J.M. 1921. *A treatise on probability.* London: Macmillan.

Kim, H., Glaeser, E. L., Hillis, A., Kominers, S. D., & Luca, M. 2023. Decision authority and the returns to algorithms. *Strategic Management Journal.*

Knill, D. C., & Pouget, A. 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27(12), 712-719.

Korteling, J. H., van de Boer-Visschedijk, G. C., Blankendaal, R. A., Boonekamp, R. C., & Eikelboom, A. R. 2021. Human-versus artificial intelligence. *Frontiers in Artificial Intelligence*, 4, 622364.

Kotseruba, I., & Tsotsos, J. K. 2020. 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*(1), 17-94.

Kralik, J. D., Lee, J. H., Rosenbloom, P. S., Jackson Jr, P. C., Epstein, S. L., Romero, O. J., & McGreggor, K. 2018. Metacognition for a common model of cognition. *Procedia Computer Science*, *145*, 730-739.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.

Kruglanski, A. W., Jasko, K., & Friston, K. 2020. All thinking is 'wishful' thinking. *Trends in Cognitive Sciences*, 24(6), 413-424.

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

Kvam, P. D., & Pleskac, T. J. 2016. Strength and weight: The determinants of choice and confidence. *Cognition*, 152, 170-180.

Laird, J. E., Newell, A., & Rosenbloom, P. S. 1987. Soar: An architecture for general intelligence. *Artificial Intelligence*, *33*(1), 1-64.

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. 2017. A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, *38*(4), 13-26.

Lake, B. M., & Baroni, M. 2023. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623 (7985), 115-121.

Lakhotia, K., Kharitonov, E., Hsu, W. N., Adi, Y., Polyak, A., Bolte, B., & Dupoux, E. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9, 1336-1354.

Lansdell, B. J., & Kording, K. P. 2019. Towards learning-to-learn. *Current Opinion in Behavioral Sciences*, 29, 45-50.

LeConte, J. 1888. The problem of a flying machine. *Science Monthly* 34, 69-77

LeCun, Y. 2017. A path to AI. *Future of Life Institute.*

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

Legg, S., & Hutter, M. 2007. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, *17*, 391-444.

Levin, M. 2024. AI: A bridge toward diverse intelligence and humanity's future. *Tufts University Working Paper.*

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., & Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.

Macmillan-Scott, O., & Musolesi, M. 2024. (Ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6), 240255.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Manning, B. S., Zhu, K., & Horton, J. J. 2024. Automated social science: Language models as scientist and subjects (No. w32381). *National Bureau of Economic Research*.

Marr, D. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Mastrogiorgio, A., Felin, T., Kauffman, S., & Mastrogiorgio, M. 2022. More thumbs than rules: is rationality an exaptation? *Frontiers in Psychology*, 13, 805743.

McAfee, A., & Brynjolfsson, E. 2012. Big data: the management revolution. *Harvard Business Review*, 90(10), 60-68.

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. 2007. A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12-12.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. 2023. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*.

McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., & Celikyilmaz, A. 2023. How much do language models copy from their training data? Evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics*, 11, 652-670.

McCullough, D. 2015. *The Wright Brothers.* Simon and Schuster.

McIntosh, T. R., Susnjak, T., Liu, T., Watters, P., & Halgamuge, M. N. 2023. From google gemini to openai q*(q-star): A survey of reshaping the generative artificial intelligence (ai) research landscape. *arXiv preprint arXiv:2312.10868*.

McCorduck, P. 2004. *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. CRC Press.

McClelland, J. L., & Rumelhart, D. E. 1981. An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review,* 88(5), 375.

McCulloch, W.S. & Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5, 115-133.

Melville, G. W. 1901. The engineer and the problem of aerial navigation. *North American Review*, 173(541), 820-831.

Merdes, C., Von Sydow, M., & Hahn, U. 2021. Formal models of source reliability. *Synthese,* 198, 5773-5801.

Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. 2024. Large Language Models: A Survey. *arXiv preprint arXiv:2402.06196*.

Morris, M. R., Sohldickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., & Legg, S. 2023. Levels of AGI: Operationalizing progress on the path to AGI. *arXiv preprint arXiv:2311.02462*.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., & Mian, A. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Newcomb, S. 1901. Is the airship coming? *McClure's Magazine*, 17, 432-435.

Newell, A., Shaw, J. C., & Simon, H. A. 1959. Report on a general problem solving program. In *International Conference on Information Processing.*

Newell, A. 1990. *Unified theories of cognition*. Harvard University Press.

New York Times, 1903 (October 9). Flying machines which do not fly.

Novelli, E., & Spina, C. 2022. When do entrepreneurs benefit from acting like scientists? A field experiment in the UK. *SSRN Working paper.*

Parr, T., & Friston, K. J. 2017. Working memory, attention, and salience in active inference. *Scientific Reports*, 7(1), 14678.

Pearl, J., & Mackenzie, D. 2018. *The book of why: the new science of cause and effect.* New York: Basic Books.

Peirce CS. 1957. The logic of abduction. In *Peirce's Essays in the Philosophy of Science*, Thomas V (ed). Liberal Arts Press: New York; 195–205.

Perconti, P., & Plebe, A. 2020. Deep learning and cognitive science. *Cognition*, *203*, 104365.

Pezzulo, G., Parr, T., & Friston, K. 2024. Active inference as a theory of sentient behavior. *Biological Psychology*, 108741.

Park, P. S., & Tegmark, M. 2023. Divide-and-conquer dynamics in AI-driven disempowerment. *arXiv preprint arXiv:2310.06009*.

Pilgrim, C., Sanborn, A., Malthouse, E., & Hills, T. T. 2024. Confirmation bias emerges from an approximation to Bayesian reasoning. *Cognition*, 245, 105693.

Pinker, S. 1994. *The language instinct: How the mind creates language.* William Morrow & Co.

Pinker, S. 2022. *Rationality: What it is, why it seems scarce, why it matters.* Penguin.

Polanyi, M. 1958. *Personal knowledge*. Routledge.

Polanyi, M. 1974. Genius in science. Cohen RS, Wartofsky MW, eds. *Methodological and Historical Essays in the Natural and Social Science*. Boston Studies in the Philosophy of Science, Vol. 14 (Springer, Dordrecht), 57–71.

Poldrack, R. A. 2021. The physics of representation. *Synthese*, 199(1-2), 1307-1325.

Popper, K., 1991. *All life is problem solving*. Routledge.

Puranam, P., Stieglitz, N., Osman, M., & Pillutla, M. M. 2015. Modelling bounded rationality in organizations: Progress and prospects. *Academy of Management Annals*, 9(1), 337-392.

Raisch, S., & Fomina, K. 2023. Combining human and artificial intelligence: Hybrid problem-solving in organizations. *Academy of Management Review*.

Ramsey, F. P. 1929. *General Propositions and Causality,* in his Philosophical Papers, ed. D. H. Mellor. Cambridge: Cambridge University Press, 1990, pp. 145–63.

Ramsey, F. P. 1931. *The foundations of mathematics and other logical essays*. Cambridge University Press.

Resnik, P. 2024. Large Language Models are Biased Because They Are Large Language Models. arXiv preprint arXiv:2406.13138.

Rescorla, M. 2015. The computational theory of mind. *Stanford Encyclopedia of Philosophy*.

Riedl, R. 1984. *Biology of knowledge: The evolutionary basis of reason*. New York: Wiley.

Roli, A., Jaeger, J. and Kauffman, S.A. 2022. How organisms come to know the world: Fundamental limits on artificial general intelligence. *Frontiers in Ecology and Evolution*, 1035-1050.

Rosenblatt, F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386.

Rosenblatt, F. 1962. *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Washington, DC: Spartan Books.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence,* 1(5), 206-215.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. 1986. Sequential thought processes in PDP models. *Parallel distributed processing: explorations in the microstructures of cognition*, 2, 3-57.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986. Learning representations by back-propagating errors. *Nature*, *323*(6088), 533-536.

Russell, S. J., & Norvig, P. 2022. *Artificial intelligence a modern approach*. London: Pearson.

Scheffer, M., Borsboom, D., Nieuwenhuis, S., & Westley, F. 2022. Belief traps: Tackling the inertia of harmful beliefs. *Proceedings of the National Academy of Sciences*, 119(32), e2203149119.

Schwartenbeck, P., FitzGerald, T., Dolan, R., & Friston, K. 2013. Exploration, novelty, surprise, and free energy minimization. *Frontiers in Psychology*, 4, 710.

Schwöbel, S., Kiebel, S., & Marković, D. 2018. Active inference, belief propagation, and the bethe approximation. *Neural Computation,* 30(9), 2530-2567.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell System Technical Journal,* 27(3), 379-423.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538.

Simon, H. A. 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 99-118.

Simon, H. A. 1965. *The shape of automation for men and management.* New York: Harper & Row.

Simon, H. A. 1980. Cognitive science: The newest science of the artificial. *Cognitive Science*, *4*(1), 33-46.

Simon, H. 1985. Human nature in politics: The dialogue of psychology with political science. *American Political Science Review*, 79, 293-304.

Simon, H. 1985. Artificial intelligence: Current status and future potential. *National Research Council Report – Office of Naval Research.*

Simon, H. A. 1990. Invariants of human behavior. *Annual Review of Psychology*, *41*(1), 1-20.

Simon, H. A. 1996. *The sciences of the artificial.* MIT Press.

Simon, H. A., & Newell, A. 1958. Heuristic problem solving: The next advance in operations research. *Operations Research*, 6(1), 1-10.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. 1992. Origins of knowledge. *Psychological Review*, *99*(4), 605.

Srećković, S., Berber, A., & Filipović, N. 2022. The automated Laplacean demon: How ML challenges our views on prediction and explanation. *Minds and Machines*, 1-25.

Sun, R. (Ed.). 2023. *The Cambridge Handbook of Computational Cognitive Sciences*. Cambridge University Press.

Tannen, D. 2007. *Talking voices: Repetition, dialogue, and imagery in conversational discourse*. Cambridge University Press.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. Science, 331(6022), 1279-1285.

Tervo, D. G. R., Tenenbaum, J. B., & Gershman, S. J. 2016. Toward the neural implementation of structure learning. *Current Opinion in Neurobiology,* 37, 99-105.

Tinbergen, N., 1963. On aims and methods of ethology. *Zeitschrift für Tierpsychologie*, *20*(4), 410-433.

Turing, A. M. 1948. Intelligent Machinery, in *Mechanical Intelligence, Collected Works of A. M. Turing.* D. C. Ince, ed. North Holland, 1992, p. 107 -127.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind*, *59* (236), 433-460.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wright, O., & Wright, W. 1881-1940. *Wilbur and Orville Wright papers.* Library of Congress. Available online at https://www.loc.gov/collections/wilbur-and-orville-wright-papers/about-this-collection/

Wuebker, R., Zenger, T., & Felin, T. 2023. The theory-based view: Entrepreneurial microfoundations, resources, and choices. *Strategic Management Journal.*

Yanai, I., & Lercher, M. 2020. A hypothesis is a liability. *Genome Biology*, *21*(1), 1-5.

Yin, H. 2020. The crisis in neuroscience. In *The interdisciplinary handbook of perceptual control theory* (pp. 23-48). Academic Press.

Yiu, E., Kosoy, E., & Gopnik, A. 2023. Transmission versus truth, imitation versus innovation: What children can do that large language and language-and-vision models cannot (yet). *Perspectives on Psychological Science*, 17456916231201401.

Zellweger, T., & Zenger, T. 2022. Entrepreneurs as scientists: A pragmatist alternative to the creation-discovery debate. *Academy of Management Review*, 47(4), 696-699.

Zhou, H. Y., Yu, Y., Wang, C., Zhang, S., Gao, Y., Pan, J., & Li, W. 2023. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nature Biomedical Engineering,* 1-13.

Zhou, P., Madaan, A., Potharaju, S. P., Gupta, A., McKee, K. R., Holtzman, A., & Faruqui, M. 2023. How far are large language models from agents with theory-of-mind? *arXiv preprint arXiv:2310.03051*.

Zhu, Z., & Kuchibhotla, K. V. 2024. Performance errors during rodent learning reflect a dynamic choice strategy. *Current Biology*, 34(10), 2107-2117.

Zhu, J. Q., & Griffiths, T. L. 2024. Eliciting the priors of large language models using iterated in-context learning. *arXiv preprint arXiv:2406.01860*.