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

Preprint · February 2024

CITATIONS
0

READS
1,452

2 authors, including:

Teppo Felin
University of Oxford
118 PUBLICATIONS **9,334** CITATIONS

SEE PROFILE

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

Teppo Felin
Utah State University
& Oxford University

Matthias Holweg
Oxford University

† An early version of this paper was presented at the *Strategy Science* "theory-based view" conference at Bocconi University (October 2023). We appreciate thoughtful feedback and comments from many participants and audience members.

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 other cognitive tasks that involve high-level reasoning and thinking. Many scholars argue that—due to human bias and bounded rationality—humans should (or will soon) be replaced by AI in situations involving high-level cognition and strategic decision making. We disagree. In this paper we first trace the historical origins of the idea of artificial intelligence and human cognition as a form of computation and information processing. We highlight problems with the analogy between computers and human minds as input-output devices, using large language models as an example. Human cognition—in important instances—is better conceptualized as a form of theorizing rather than data processing, prediction, or even Bayesian updating. Our argument, when it comes to cognition, is that *AI's data-based prediction is different from human theory-based causal logic*. We introduce the idea of belief-data (a)symmetries to highlight the difference between AI and human cognition, and use "heavier-than-air flight" as an example of our arguments. Theories provide a mechanism for identifying *new* data and evidence, a way of "intervening" in the world, experimenting, and problem solving. We conclude with a discussion of the implications of our arguments for strategic decision making, including the role that human-AI hybrids might play in this process.

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 strategic games, standardized tests, and cognitive tasks that involve high-level thinking and strategic reasoning. For example, various AI engines beat humans not just in chess—which for decades served as a key benchmark of AI capability (Bory, 2019; Simon, 1985)—but also in games like Jeopardy. AI engines also play multiplayer strategy games like Diplomacy at an extremely high level—games that involve sophisticated negotiation, complex interaction with others, alliances, deception, and understanding other players' intentions (Kramar et al., 2022). The latest versions of AI also now outperform 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 doctors in diagnosing many illnesses (e.g., Zhou et al., 2023). AI scholars argue that even the most human of traits, 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 high-level reasoning and cognitive decision making (e.g., Christian and Griffiths, 2016; Grace et al., 2024, Legg and Hutter, 2007; Morris et al., 2023). One of the modern pioneers of AI argues that large language models already are sentient and intelligent, and that "digital intelligence" will inevitably surpass human "biological intelligence"—if it has not already done so (Hinton, 2023; also see Bengio et al., 2023).

Compared to machines, the cognitive and computational limitations of humans seem obvious. Humans are biased and boundedly rational (for a review, see Chater et al., 2018; also see Kahneman, 2003; Kahneman, 2011). Humans are selective in what data they attend to or sample. 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). 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—which provides the broad basis of AI—is that they can handle vast amounts of data and process it in powerful ways.

In this paper we offer a contrarian view of AI relative to human cognition. We first revisit the historical origins of the claim that equates computation and information processing with human cognition. AI builds on the idea that cognition is a generalized form of information processing, an input-output device. To illustrate differences between human and computational processing, 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 theoretically "top-down" rather than "bottom-up" from data. We introduce the notion of data-belief (a)symmetry and the role this respectively plays in AI and human cognition. We use "heavier-than-air" flight as an example to illustrate our point. Human cognition is forward-looking, necessitating data-belief asymmetries which are manifest in theories. 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* data, observations, and experimentation. We conclude with a discussion of the implications for these arguments for strategic decision making, along with briefly highlighting opportunities for considering human-AI hybrid systems.

Before proceeding, we need to briefly comment on the title of this paper—"theory is all you need." Our title of course echoes the title of the "attention is all you need" article that (among others) 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 key—often unrecognized—aspect of human cognition, one that is not easily replicable by machines and AI. We emphasize the role of theory in human cognition, particularly the ways in which humans counterfactually think about and practically "intervene" in the world. This differs radically from AI-based models that are theory-free, or that place

primacy on data and prediction. While prediction-based approaches are important, our emphasis instead is on how human cognition enables the development of forward-looking theory-based causal logic, experimentation, and problem solving. Theory-based causal logic allows decision makers to go *beyond* data and prediction, enabling the generation of *new* data.

AI = MIND: A REVIEW OF COGNITION AS COMPUTATION

Modeling the human mind—thinking, rationality, and cognition—has been the central aspiration and ambition behind artificial intelligence (AI) from the 1950s to the present (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),[1] the Bayesian brain and predictive coding or processing (Friston and Kiebel, 2009; Hohwy, 2013, 2020).

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 seen as a type of generalized and comparable input-output device, where inputs—such as stimuli and cues, or "data"—are processed to yield varied types of outputs—such as 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 (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

[1] As Johnson-Laird argues, any "any scientific theory of the mind has to treat it as an automaton" (1983: 477).

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).

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 develop a way 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 Simon—in 1958, in the journal *Operations Research*—"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, as 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—that is, to mimic or even exceed human capability—their application domains have turned out to be extremely specific (in terms of the problems they actually solve).

Of necessity, the problems that these AI models focused on were relatively delimited, often using games as a context, with the hope that insights from these models could be generalized once the focal problem had been solved. For example, AI pioneer Herbert Simon revisited a specific set of cognitive problems and tasks throughout his long career. His favorite practical contexts for studying and modeling cognition were (a) chess, which he called the "microcosm" of cognition and the "fruit fly of artificial intelligence" (Simon, 1985), (b) the Tower of Hanoi puzzle, (c) the missionaries and cannibals problem, or (d) various cryptarithmetic problems (problems where letters are replaced by numbers). Other scholars focused on problems related to memory, including how humans might

parse sentences, chunk items or recall a set of stimuli (e.g., Anderson, 1976; Miller, 1956). While all of these tasks are "cognitive" in some fashion—and in principle amenable to different forms of computational modeling—they did not meaningfully attain any form of generality. Naturally the types of problems that any given cognitive model or architecture is trained on, or motivated by, has implications for which variables are seen as central, thus limiting overall generality.

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 the idea of "computational rationality" (Gershman et al., 2015).

Thus, 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., 2021). 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 any given model is highly focused on 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 (e.g., 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." 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.

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"—learning directly from the data. Vast inputs and raw data are fed to these systems to recognize correlations and associations, or in short, patterns. The weakness of the aforementioned symbolic systems is that that approach is 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, and particularly, perception. Rosenblatt (1958, 1962; building on Hebb, 1949) proposed the earliest form of a neural network in his model of a "perceptron," the functional equivalent of an artificial neuron. Rosenblatt's work on the perceptron aimed to replicate the human neuron, which 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, 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).[2] 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. This feedback-based learning process was a precursor to the concept of Hebbian learning (Hebb, 1949), which posits the (now) relatively cliché idea that the "neurons that fire together wire together."

In the intervening decades, research on artificial neural networks has progressed radically from simple linear classifiers to highly complex, multilayer non-linear models capable of sophisticated feature learning and pattern recognition through weighting and updates based on extremely large datasets (e.g., Aggarwal, 2018; Shazeer et al., 2017). Various forms of machine learning—for example: supervised, unsupervised, reinforcement—have enabled radical breakthroughs in image recognition and computer vision, recommender systems, game play, text generation and so forth. And 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), and inductive reasoning by humans and AI (Bhatia, 2023), learning to make inferences (Dasgupta et al., 2020). All of these models of learning build on neural networks in various forms.

Now, our overall purpose is *not* to review the specific, exhaustive details of these models (excellent reviews can be found elsewhere, e.g., Aggarwal, 2018; Russell and Norvig, 2022). Rather, in the above we have sought 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. Cognition and AI are said to be deeply connected. The underlying premise of this work, by way of summary, 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.

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

We disagree with this premise, 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 simply want to highlight how the analogy between AI and humans breaks down when it comes to understanding the mind and cognition—with important derivative consequences for how we think about judgment and decision making. 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 "so that it can be a true 'window into the mind' "(Chomsky, 2020: 321; also see Pinker, 1994). Language, mind and cognition are inextricably linked and thus language provides an important "test" and context for understanding human and artificial intelligence. Furthermore, large language models have already been argued to be sentient, with some arguing that they already closely mirror 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). Turing argued that 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" (Turing, 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 (e.g., chess, poker), cryptography, mathematics—and argues that "learning of languages would be the most

impressive, since it is the most human of these activities" (Turing, 1948: 117). The metaphor of the infant shows up in modern applications of machine learning (e.g., Zaadnoordijk et al., 2022), and language (in the form next-word prediction) of course is one of the more impressive feats of modern AI.

How Machines Learn Language

To illustrate this process of machine learning, next we carefully consider modern large language models (LLMs), in the input-output form laid out by Turing. 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 associational interactions amongst that data. In the context of an LLM, the training data is 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 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 years to read 13 trillion words in the 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 "transformer" architecture (Vaswani et al, 2017) builds on foundational work in the domain of artificial neural networks, touched on by us previously. Artificial neural networks have turned out to be extremely general and applicable not just to text but varied domains associated with image recognition and computer vision, including multi-modal applications that combine various types of data.

In terms of the training of a large language model, the tokenized words are submitted for algorithmic processing based on a predetermined sequence or input length. That is, 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 LLMs models use input lengths

of 2,048. To illustrate this, if a large classic novel, like (say) Dostoyevsky's *Brothers Karamazov*, is included in a pretraining dataset, the novel's roughly 400,000 or so word/tokens would represent 195 or so unique, independent training datapoints or training examples that are fed into the system—given the predefined token sequence of 2,048. So, a 13 trillion token training dataset, parsed into 2,048 length sequences (also called "context window or length"), would contain 6.3 billion unique sequences from which the algorithm can learn.[3]

From this vast data—through pretraining—the LLM learns associations and correlations between various elements of language: words relative to each other, their relationships, ordering, frequencies, and so forth. These 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. For instance, it discerns that certain words are more likely to appear in specific contexts. 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 Minaee et al., 2024; Naveed et al., 2023; Zhao et al., 2023). The end-result of this training is an AI model that is capable of language: that 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 first 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 point out differences, particularly as these provide the foundation for our subsequent arguments related to human 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;

[3] Sequence length is important because it allows for the LLM to understand context. Newer models have introduced sequence lengths in excess of one million tokens.

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.[4] The inputs differ radically in terms of quantity (sheer amount), but also in terms of their quality.

Namely, the spoken language that an infant 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 and planned, edited and polished, more dense, featuring more complex vocabulary (e.g., Halliday, 1989; Tannen, 2007). More importantly, the functional purposes and uses of spoken versus written language differ significantly. Spoken language is immediate, interactive, focused on coordinating, expressing, and practically doing things. While written language also features these elements, the purpose of written language tends to be more on the communication of complex information. The vast bulk of the 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., Gutenberg), lyrics, blog posts, news articles, and so forth. This data is far "cleaner," far more correct grammatically, and organized. In a statistical sense, it contains less "noise" and thus offers greater predictive power. Even the vast stores of Wikipedia articles that are included in most LLM pretraining datasets are the end result of thousands of edits to ensure readability, accuracy, and flow.

Clearly humans learn language under different conditions and with different types of inputs.[5] In fact, it could readily be argued that the human capacity for language develops qualitatively and quantitatively differently from how machines learn language. Humans (somehow) learn language from extremely sparse, impoverished, and highly unsystematic inputs and data. This logic is aptly captured by Chomsky:

[4] For an infant to be exposed to the same 13 trillion tokens represented by the pretraining of some current LLMs, it would take 650 million days or roughly 1.8 million years.

[5] Human learning of language (or anything else) of course is multimodal. Thus it is hard to separate linguistic inputs and learning from other forms of input and learning that are tightly coupled (like visual, auditory, and other forms of input and cues that accompany language). Thus our emphasis on how humans learn from comparatively sparse data (when only accounting for linguistic inputs) does not account for that vast "pretraining" that is included through other modalities. As discussed below, we nonetheless think the simple focus on inputs and outputs fundamentally misses what is central about human cognition (cf. Felin and Koenderink, 2021).

> 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*).

So, beyond the quantitative and qualitative differences in inputs, 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"). LLMs are generative in the sense that they are able to create novel outputs by probabilistically drawing on (or sampling from) the vast combinatorial possibilities in the associational and correlational network of word orderings, embeddings, frequencies, co-occurrences, and contexts encountered in the training data (Vaswani, 2017).[6] The text LLMs produce 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 input data probabilistically have on the output. For example, in a sentence completion task, the weights 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 not only have compositional novelty (in terms of novel ways of saying the same thing), but they also manifest analogical generalization (McCoy et al., 2023). Any recognition 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 a next-word prediction engine, LLMs do a tremendous job.

From Representation to New Knowledge?

Now, a central question is, is the output generated by an LLM "intelligent?" Certainly LLMs seem to manifest the sparks of intelligence, that is, their output appears to be intelligent. But this seeming intelligence is

[6] 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).

partly an epiphenomenon of the *fact that the same thing can be stated, said, and represented in indefinite ways*. To offer a visual metaphor, we might think of a given output of an LLM as a map that changes how it represents a specific territory—pulling from varied styles, structures, and color codings across other maps. There are highly varied ways of representing the same territory, just as there are highly varied ways of saying the same thing. This point is worth emphasizing. Language inherently allows for vast numbers of ways of communicating, rephrasing, paraphrasing or summarizing any given idea, fact, or sentiment. LLMs' ability to generate this type of isomorphism—essentially, different ways of capturing the same information—is crucial for their performance. An LLM learns different patterns and meta-patterns for representing something. Furthermore, the ability to say something in many ways allows LLMs to adapt their responses to the specific preferences, prompts, and requirements of a task or user.

It is important to note that the revolutionary breakthrough that gave rise to LLMs—the transformer architecture—was specifically developed in the context of language *translation* (see Vaswani et al., 2017). LLMs can be seen as "translation generalized." That is, LLMs are a 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 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. Next word prediction—using the weights and parameters found in vast training datasets and the process of probabilistically drawing from this training—turns out to enable surprisingly rich combinatorial and sophisticated outputs. The learning of the LLM is embodied in the relationships found between words which are sampled to enable stochastic generativity.

To highlight further (metaphorical) similarities between AI-based neural networks and cognition, consider large language models and cognitive processes like predictive processing (Pezzulo, Parr and Friston, 2024: a rough synonym for active inference, the free energy principle, the Bayesian brain, and predictive coding). Both large language models and predictive processing seek to engage in error minimization and iterative optimization, where a system is essentially navigating a high-dimensional space to find a state that minimizes error and surprise. LLMs learn from the training data and predictive processing learns from its environment (Hohwy, 2020). LLMs aim to

reduce the difference between their predictions (e.g., 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 their 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). Back-propagation, a fundamental mechanism in training neural networks, and the concept of error minimization in predictive processing (Friston et al., 2009) share a conceptual similarity in that both involve iterative adjustments to minimize some form of error or discrepancy. 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.

But can an LLM—or any prediction-oriented cognitive AI—truly generate some form of new knowledge? We do not believe they can.[7] One way to think about this is that an LLM could be said to have "Wiki-level knowledge" on varied topics in the sense that these forms of AI can summarize, represent, and mirror the words (and associated ideas) it has encountered in myriad different and new ways. On any given topic (if sufficiently represented in the training data), an LLM can 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, so an LLM is highly unlikely to somehow bootstrap knowledge beyond the combinatorial possibilities of the data and word associations it has encountered in the past. Here our goal is not to dismiss, at all, the remarkable feats of LLMs, nor other applications of machine learning. The fact that an LLM can readily 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 knowledge—or more specifically, its linguistic and associational word-structure—in its training data, and it is able to summarize it in coherent, diverse, and fluent ways. Thus the idea of LLMs as "stochastic parrots" or glorified auto-complete (Bender et al., 2021) underestimates their ability. Equally, ascribing an LLM the ability to generate new knowledge overestimates its ability.

[7] Though we of course recognize that there is significant disagreement on this point (e.g., Franceschelli and Musolesi, 2023).

We might concur that LLMs are powerful and creative "imitation engines" (in stochastically assembling words)—though not linguistically innovative compared to children (Yiu, Kosoy, and Gopnik, 2023). But 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 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 and many others. Ptolemy's careful trigonometric and geometric calculations, along with the associated astronomic observations, would offer further evidence for the geocentric view. The texts would feature word associations that highlight how the motions and movements of the planets could be predicted with remarkable accuracy based on a geocentric view. In short, the evidence—as inferred from the word associations found in the training data—would overwhelmingly be against Galileo. LLMs do not have any way of accessing truth, beyond mirroring and restating what is found in the text.

Notice that even if alternative or "heretical" views were included in the training data (like the work of Copernicus, even though his work was banned), the logic of this work would be dwarfed by all the texts and materials that supported the predominant paradigm, a geocentric view.[8] In other words, the overwhelming corpus of thousands of years of geocentric texts would vastly "outweigh" Galileo's view in the training data. The frequency with which the geocentric view has been mentioned, summarized, and discussed in the training data

[8] 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 censored (Gingerich and Maclachlan, 2005).

necessarily imprints itself onto the output of the LLM. As the LLM has no grounding in truth—beyond what can be inferred from text—it would say that Galileo's view is a fringe perspective, a delusional belief.

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—*an AI trained with past data has no way of doing so, no way of being forward-looking*. For example, the eminent astronomer Tyco Brahe made and famously published claims on 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 among the educated elite. An LLM would have no way of arbitrating between Copernicus's seeming delusions about heliocentrism nor Brahe's astrological writings. The LLM can only represent and mirror the predominant and existing conceptions—in this case, support for the geocentric view of the universe—it finds in the statistical association of words in its training data. It is important to recognize that the way an LLM gets at "truth" is by finding *more frequent* mentions of a true claim (in the form of statistical associations between words)—and less frequent mentions of a false claim. *For an AI trained with text, truth is an epiphenomenon of the fact that true claims have been made more frequently than false claims in the training data*. Put differently, *truth emerges as a byproduct* of these statistical patterns and frequencies rather than from the LLM developing an intrinsic understanding of—or ability to bootstrap—what is true or false in reality.

Notice that some LLMs have sought to engineer around this problem 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 a form of expertise (Du et al., 2023; Shazeer et al., 2017). Furthermore, ensemble approaches—which combine 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 these experts have said, rather than any form of forward-looking projection or causal logic.

Thus it is important to keep in mind that the inputs of any LLM are human inputs, and they roughly represent what we know so far (in the form of word associations and correlations). Inherently they cannot go beyond the realms covered in their training data. There is no mechanism for somehow bootstrapping forward-

looking beliefs and knowledge beyond what can be inferred from the existing statistical associations and correlations found in the training data itself.

The Primacy of Data and Data-Belief Symmetry

So far, the central problem we have highlighted is that learning by machines 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—struggles with generating new knowledge, postulating beyond what it has encountered in its pretraining data. Next we extend this problem to the more general emphasis on the primacy 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. 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 (or any large corpus or environment) has a particular statistical and physical structure, and the goal of a system is to accurately learn from it and reflect it. As summarized 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*). 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.[9] 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).[10]

[9] The problems with this approach have not just been discussed by us. For related points in neuroscience, see Yin, 2020.
[10] 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 and model architecture, but also loss functions and metrics, regularization and generalization techniques, valued outputs, and so forth.

Bayesian cognition also posits that learning by humans and machines can be understood in terms of probabilistic reasoning, akin to Bayesian statistical methods (Griffiths et al., 2010). This framework conceptualizes sensory inputs, perceptions and experiential evidence as data, which are continuously acquired from the environment. The cognitive process involves sampling from a probability distribution of possible states or outcomes, informed by this data. Crucially, this model emphasizes 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 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. As suggested by Bayesian models, we believe things to the extent to which we have data and evidence (Pinker, 2021). Beliefs therefore 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).[11] Knowledge can be seen as the accumulated inputs, the data and evidence that make up our beliefs. And *the strength* or *degree* of any belief should be commensurate 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. This approach focuses 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—the weight of the evidence. A rational decision maker will form their beliefs about any given thing by taking into account the available data and evidence (Pinker, 2021).

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

But what about "edge cases?" That is, what about situations where an agent correctly takes in all the data and evidence, but yet somehow turns out to be wrong? Perfect or omniscient information processing models do not offer a mechanism for explaining change, or situations where data and evidence-based reasoning might lead to poor outcomes (cf. Chater et al., 2018). While learning-based models of knowledge enable "belief updating"—based on new data—there is no mechanism for explaining *where new data comes from*. And what if the data and evidence are contested? Overall, the problem is that an agent could be a perfectly rational, Bayesian decision maker—accurately accounting for all the relevant data and updating accordingly—and yet be wrong.

These situations are problematic for input-output models 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 *asymmetry* is essential for the generation of new knowledge. The existing literature in the cognitive sciences and AI has focused on one side of the data-belief asymmetry, namely its *downside*—or, the negative aspects of data-belief asymmetry. This downside includes all the ways in which humans persist in believing something *despite* clear evidence to the contrary. This includes a large literature which focuses on human biases in information processing—the suboptimal and biased ways that humans process and use data (Chater, 2018; Kahneman, 2011; Kahneman et al., 2021). This is readily evident in the vast literatures that focus on various data-related pathologies and biases, including confirmation bias, selective perception and sampling, and availability bias. All of these biases are the result of distorted and delusional beliefs (Pinker, 2021).

But what about the *positive side* of data-belief asymmetry? What about situations where beliefs appear delusional and distorted—that is, contrary to established evidence and facts—but turn out to be correct? Here we are specifically talking about beliefs that may outstrip, ignore, and go beyond existing data. Notice, as we will discuss, this perspective is not completely unchecked or untethered from the world. Rather, this form of data-belief asymmetry is forward-looking, and can enable the identification of new data and evidence, 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 to the early 1900s. We introduce this example here, though 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, after all, can readily be defined as belief contrary to evidence and reality: 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 pointed to birds as a form of "positive" evidence for the belief that humans might also fly. This was a common argument.[12] But the idea that bird flight somehow provided "hope" and evidence for the plausibility of human flight was seen as delusion and put to 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 one of the critical determinants of flight. To further reinforce this point, LeConte (and other scientists) noted the ubiquity of insects flying as well. But LeConte's central point (and evidence) was that clearly no bird above the weight of 50 pounds is able to fly, and 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—who was one of the foremost astronomers and mathematicians of his time—added in 1901 that "the most numerous fliers are little

[12] 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).

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."

LeConte's attempt to use evidence related to birds (and their weight) to disprove the possibility of human powered flight highlights a problem with data. It is hard to know what data and evidence might be *relevant* for a given belief and argument. The problem is, as succinctly put by Polanyi, that "things are not labeled *evidence* in nature" (1957: 31). What is the relevant data in this context? Did flight have something to do with feathers or with wings or with evolution or with the weight of birds? If it had something to do with wings, was it wing shape, wing size, or wing weight?[13] In short, it is extremely hard to know what data might be relevant for the formation of a particular belief.

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. More often than not, for our evidence we 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 more recently begun to recognize this. Bayesian and other probabilistic models of cognition have also introduced the idea of "the reliability of the source" when considering 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 important.

This is readily illustrated by our discussion of heavier-than-air flight. So, what might happen if we update our beliefs about the plausibility of human flight by focusing on reliable sources—like the evidence and reasoning of science and scientists? In most instances, this is a good strategy. But updating our belief on this basis, when it comes to heavier-than-air flight, would lead to the conclusion that human powered flight was delusional and impossible. Again, LeConte argued that flight was impossible using evidence. But not only should we update our belief based on this evidence, but we should also update our belief based on the fact that he was a prominent scientist; in fact, he became the eventual President of the leading scientific association at the time, the American

[13] Notice that 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.

Association for the Advancement of Science. And LeConte was not alone. He was part of a much broader scientific consensus at the time about the impossibility of human flight. 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. The aforementioned mathematician-astronomer Simon Newcomb (1901) also argued in the early 1900s—in his article, "Is the airship coming?"—that 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 contrary data, evidence, and scientific consensus—hold onto the belief that human flight is possible? The data and evidence against the plausibility of flight was overwhelming. No rational Bayesian should believe in flight. The evidence was not just empirical (in the form of LeConte's bird and other data) and scientific (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—Samuel Langley spectacularly failed 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 objective data, evidence, and scientific consensus—turned out to eventually be correct. However, we think this example offers an instance of a more generalizable process, where humans intervene in the world and "go beyond the data." Cognitive psychologists focused on beliefs and belief updating generally engage in the opposite exercise related to data and beliefs, where a belief objectively should be updated by data, but it is not (Festinger et al., 1956)—as illustrated by conspiracy theories and other forms of biased data sampling (Chater, 2018; Kahneman, 2011; Pinker, 2021).

Heterogeneity in beliefs is seen as problematic. And in many instances, it is. But heterogeneity in beliefs—belief asymmetry—is also the lifeblood of new ideas, new forms of experimentation, and new knowledge, as we discuss next. Our goal here is simply to highlight that this positive form of data-belief asymmetry also deserves consideration, if we are to account for cognition and how human actors counterfactually think about, intervene in, and shape their surroundings. Computational models of information processing offer an omniscient ideal that turns out to be useful in many instances, though not in all instances.

THEORY-BASED CAUSAL LOGIC

Building on the aforementioned data-belief asymmetry, next we discuss the cognitive and practical process by which humans engage in forward-looking and generative theorizing that enables them to, in essence, go "beyond the data" (or more specifically, go beyond existing data to find new data). We specifically highlight how this form of cognitive and practical activity differs from computational and information processing-oriented forms of cognition, and how it enables 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* (Pearl and Mackenzie, 2018), to develop a causal logic. We extend the aforementioned 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 process of what we call "theory-based causal logic."

Our foundational starting point—building on Felin and Zenger (2017)—is that cognitive activity is theoretical or scientific activity.[14] 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

[14] 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 in their environments. 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; also see Uexküll, 1926). 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.

scarcely account for the vast outputs—thus pointing to a human generative capacity to theorize (cf. Chomsky, 1975).

Importantly, cognition and language are used to *do* things. This is also the basis of the so-called "core knowledge" argument in child development (Carey and Spelke, 1996; Spelke et al., 1992). That is, humans develop knowledge like scientists, through a process of conjecture, hypothesis, and experimentation. While empiricist approaches focus on the primacy of data and environmental inputs, theory-based approaches focus on the active role of humans in not just learning about but also experimenting with and changing their surroundings (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 early article in developmental psychology titled "If you want to get ahead, get a theory" (Karmiloff-Smith and Inhelder, 1974). And this echoes Kurt Lewin's famous maxim, "there is nothing as practical as a good theory" (1943: 118; cf. Dewey, 1916). The central point here is that theories are not just for scientists, theories are pragmatically useful for anyone. 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 of strategy extends the above logic and emphasizes the importance of theorizing and theories in economic contexts, with widespread implications for cognition, decision making, and governance (Felin and Zenger, 2017). The central idea behind the theory-based view is that economic actors can (and need to) develop firm-specific theories. Theories do not attempt to map existing realities but 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" factor markets (cf. Barney, 1986), by seeing and searching the world differently, and by identifying new sources of value (Felin, Kauffman, and Zenger, 2021). Awareness for new possibilities is cognitively developed top-down (Felin and Koenderink, 2022; Koenderink, 2012). Theories also have central implications for how to efficiently organize the governance of value creation (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).[15] The practical implications of the theory-based view have also led to the development of managerial tools to enable value creation by startups, economic actors, and organizations (e.g., Felin, Gambardella, and Zenger, 2021).

Now, our goal in this section of the paper is *not* to review the theory-based view. Rather, our goal now is to further build out the cognitive aspects of the theory-based view, and to highlight how the human capacity for theorizing differs from AI-infused and computational approaches to cognition which focus on prediction. We highlight how a theory-based view of cognition enables humans to intervene in the world—beyond the given data—and not just to process or represent it. We concurrently continue to highlight how this approach *differs significantly* from the arguments and prescriptions suggested by computational and AI-inspired approaches to cognition. 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).

Belief-Data (A)symmetry Revisited

Heterogeneous beliefs provide the initiating impetus for the theory-based view. From our perspective, for beliefs to be a relevant concept for strategy, beliefs do not necessarily—in the first instance—need to be based on data. We are specifically interested in forward-looking beliefs, that is, beliefs that *presently* lack evidence or even go against existing evidence. Forward-looking beliefs, then, are more *in search of* data rather than *based on* 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, or where the corroborating data simply has not (yet) been identified, found, or experimentally generated. Beliefs do not automatically verify themselves (James, 1967)—often they require some form of action. The search for data in support of an uncommon belief necessarily looks like irrational, motivated reasoning. To 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

[15] There are close 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).

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 and 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 seeming contrarian or delusional beliefs, it is worth emphasizing the critically 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 evidence, rather than a likelihood of taking action—a significant difference.

Notice the implications of this in a context like heavier-than-air flight. Beliefs played a central role in motivating action on the part of aviation pioneers *despite* overwhelming data and evidence. 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." Wilbur clearly recognized that his belief about flight appeared delusional to others, as is evident from his many letters (Wright and Wright, 1881-1940). But this belief motivated him to experiment and problem solve and to make the seemingly delusional belief a reality (only four short years later). And 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, Kelvin quickly 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 a delusional endeavor. Without the slightest belief in the possibility of human flight, he did not work toward nor want to support anything related to it. But for the Wright brothers, the possibility of powered flight was very

much—in the words of William James (1967)—a "live hypothesis." *Despite* the data, they believed human flight might be possible, and acted accordingly.

This approach presents problems for the idea of rationality (cf. Felin, Koenderink, and Krueger, 2017). To be a rational human being, knowledge should be based on evidence. In a sense, the concept of "beliefs" is not even meaningfully needed, as one can instead just talk about knowledge. 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 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 belief-data asymmetries, of the positive type, can be highly useful. They direct our awareness toward new data and possible experiments. This seeking-data-to-verify-a-belief of course is the epitome of motivated reasoning and confirmation bias. It points to "the bad habit of seeking evidence that ratifies a belief and being incurious about evidence that might falsify it" (Pinker, 2021: 13). Holding an asymmetric belief seems to amount to "wishful thinking"—"protecting one's beliefs when confronted with new evidence" (Kruglanski, Jasko and Friston, 2020: 413). For example, the Wright brothers were constantly confronted with new, disconfirming evidence, including Samuel Langley's public failures with flight, only a few months before their own success. But in these instances, ignoring the data and evidence—*not* updating beliefs—turned out to be the correct course of action.

Now, perhaps the examples we have drawn on are rare and exceptional instances—instances that we can only discuss post hoc once the delusional or veridical nature of the beliefs in question is actually known. But notice that scientists readily point to the other side of data-belief asymmetry: where the weight of the data is irrationally ignored. Beliefs updates are not rational. Here we are pointing to the other side of the equation, where the weight of the belief is (irrationally or rationally) ignored. Again, scientists have largely focused on the negative instances of data-belief asymmetry, where humans fail to appropriately update beliefs based on data (Pinker, 2021). We argue that the positive aspects of data-belief asymmetry are also extremely important to understand, where humans *rightly* fail to update beliefs based on (existing) data, and where their seemingly delusional belief turns out to correct.

Importantly, in this context, the *degree* or *strength* of belief does not need to—as is done in extant models of cognition (e.g., Pilgrim et al., 2024; Pinker, 2021)—be directly tied to commensurate or symmetric data or evidence. Beliefs have a causal role of their own and can be measured by our propensity to act on them (Ramsey, 1931). Of course, willingness-to-act on beliefs does not assure us that they are true. But they are an important motivation for action (Ajzen, 1991; Bratman, 1987).[16] And this is not to say that data does not matter when it comes to beliefs. Rather, beliefs provide the impetus for action, for creating the experimental conditions for (eventually) finding or producing the *right* data.

From Beliefs to Directed Experimentation and Problem Solving

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 generate novel data and reveal solutions—which presently are hidden (Felin, Gambardella and Zenger, 2021). Put differently, if the current conditions of the world do not provide data to support the plausibility of a belief, agents can seek to create the conditions for finding the relevant evidence or engage in directed experimentation and problem solving to enable belief realization.

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, but rather that we counterfactually think about possibilities and future states, with an eye toward taking action and providing the *right* evidence. This shifts the locus from information processing and prediction to doing and experimentation. This involves actively questioning and manipulating causal structures, allowing for a deeper exploration of "what if" scenarios. Counterfactual thinking empowers humans to probe hypothetical alternatives and dissect causal mechanisms—to understand the all-important *why*-question—offering insights into necessary and sufficient conditions for an outcome (Felin, Gambardella, Novelli, and Zenger, 2023). This approach is significantly different from input-output and information processing-oriented models of AI and computational cognition, and even various "data-driven" approaches to decision making. These types of models focus largely on *patterns* based on associations and correlations found in the data. But they lack an ability to understand underlying causal structures,

[16] 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 possibilities, and possible interventions (Pearl and Mackenzie, 2018). This is the role of theory-based causal logic.

A doing-orientation is aptly illustrated by the difference between how data-oriented and evidence-based scientists thought about the possibility of 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.[17] 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 different data related to bird flight. Rather than look at the weight of those birds that flew and those that did not, the Wright brothers engaged in observational studies of the *mechanics* of bird flight and anatomy, for example, carefully studying the positioning of bird wings when banking and turning.

The 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 that flight was impossible. 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 (who died attempting flight in 1896). The Wright brothers notes and letters reveal that they carefully studied the flight attempts and aircraft of earlier pioneers like George Cayley, Alphonse Penaud, Octave Chanute and others (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. The result of these efforts was that the Wright brothers reasoned that if they solved three problems—lift, propulsion, and steering—they would be

[17] 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.

able to fly. To illustrate the power of developing a theory-based causal logic, and identifying a problem to solve—coupled with directed experimentation—we briefly discuss how they addressed 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 (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 specific, testing various wing shapes, sizes and angles. They quickly realized that not everything needed to be tested at scale and that this experimentation could more safely and cost effectively be done in laboratory conditions. Thus they constructed their own wind tunnels. Targeted tests within these tunnels enabled them to learn the central principles of lift. They measured everything and kept meticulous track of their data—data that they generated through experimentation. 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 central also for solving propulsion and steering or control. And more generally, the Wright brothers were careful scientists in every aspect of this process. 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 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 and experiments (Wright and Wright, 1881-1940).

The Wright brothers' approach to flight offers a useful case study of how theory-based causal logic enables belief-realization, even in the context where beliefs seemingly are not warranted by existing data or science. Based on their extensive study and theory, 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 equally 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 this enables the emergence of heterogeneity and the creation of novelty and value. Next we further explore the implications of this argument for computational forms of cognition and strategic decision making.

DISCUSSION: THE LIMITS OF PREDICTION FOR STRATEGIC DECISION MAKING

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 the nature of thinking, intelligence, and the mind. The prevailing assumption in much of the cognitive sciences is that the human mind is a computational, input-output system. This approach is aptly summarized by Christian and Griffiths (2016: 12):

> Looking through the lens of computer science can teach us about the nature of the human mind, the meaning of rationality, and the oldest question of all: how to live. Examining cognition as a means of solving the fundamentally computational problems posed by our environment can utterly change the way we think about human rationality.

This computational approach to understanding the human mind and rationality also has widespread practical implications, specifically for areas such as judgment and decision making—leading to what Christian and Griffiths (2016) call a "computer science of human decisions." There is increasing consensus that modern decision environments are overwhelmingly data intensive and therefore more suitable for algorithmic rather than human processing. As Kahneman and his coauthors put it, the problem is that humans "cannot hope to use information as efficiently as an AI model does." And "when there is a lot of data, machine-learning algorithms will do better

than humans" (Kahneman, Sibony and Sunstein, 2021: 112).[18] As a result, many argue that decision making should (or will) increasingly be relegated to AI.

In this article we have emphasized how human cognition differs from AI. However, the advancements in AI—independent of these differences—do raise important questions about decision making. In particular, what about *strategic* decisions? Will strategy also be relegated to AI? We fundamentally think that strategic decisions are not amenable (solely) to AI-based cognition or computational decision making, for many of the reasons discussed in this article. *Most importantly, AI-based models are based on backward-looking data and prediction rather than any form of forward-looking theory-based causal logic* (that is based on directed experimentation, problem solving, and new data). We claim that emphasizing or relying on prediction alone is a debilitating limitation for strategic decision making.

To further justify our claim, we briefly revisit why *prediction*—the hallmark of AI—is a problematic basis for strategic decision making. The overall logic of prediction is aptly summarized by Agrawal et al (2022). They rightly argue that, stripped down to its essence, "AI is a *prediction* technology" (2022: 22-32). A central question then is—is prediction a central ingredient of *strategic* decision making as well?

To understand this, the nature of prediction needs to be unpacked. Agrawal et al (2022) argue that "data provides the information that enables a prediction" and "predictions are a key input into our decision making." We might summarize this argument by calling it a hierarchy (of sorts), that goes from data to information to prediction to decision—or, data->information->prediction->decision.[19] This framework applies to routine decision making, but Agrawal et al also argue that "prediction is at the heart of making decisions *amid uncertainty*" (2022: 7). If we use "decision making under uncertainty" as a definition of strategy, then this puts prediction at the heart of strategy as well. The centrality of prediction is corroborated by one the pioneers of AI, Yann LeCun (2017), who argues that "prediction is the essence of intelligence."

The problem with putting prediction center stage in strategic decision making (or intelligence) is that prediction inherently is based on past data. Even vast amounts of data cannot necessarily anticipate the future. For

[18] This echoes strong claims made by others. For example, Harari argues: "In the last few decades research in areas such as neuroscience and behavioural economics allowed scientists to hack humans, and in particular to gain a much better understanding of how humans make decisions. It turned out that our choices of everything from food to mates result not from some mysterious free will, but rather from billions of neurons calculating probabilities within a split second. Vaunted 'human intuition' is in reality 'pattern recognition.'" (2018: 29).

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

various routine decisions—related to operations in an organization—prediction is a fantastic tool. Vast data can be highly powerful in situations that match the past. But note that the purpose of prediction machines and prediction-based cognition is to *minimize* surprise and error. This also matches the strong emphasis that many scholars of judgment and decision making put on "consistency"—and the eagerness to avoid noise (see Kahneman, Sibony, and Sunstein, 2021).

But the purpose of strategic decision making is—in an important sense—more about *maximizing* surprise and error (or, what might look like "error" to others) rather than minimizing it. In a strategy context, the most impactful decisions are the ones that seem to ignore (what to others looks like) objective data. In an important sense, strategic decision making has more to do with unpredictability than prediction.

However, 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). Theories reveal new data—which emerge from the process of experimentation and problem solving. This is precisely what the Wright brother's theory of flying—and associated experimentation and problem solving—illustrates. And this is just as evident in the context of the Wright brothers as it is in the strategies of companies like Airbnb. The Airbnb founders had a contrarian belief in the plausibility of using vacant apartments as hotel accommodation. While in retrospect we might point to plausible data confirming this belief, at the time the success of their venture was seen as highly unlikely by many, including sophisticated venture capital investors (Felin, Gambardella, Novelli and Zenger, 2024). But their efforts at directed experimentation and problem solving led to the creation of significant economic value.

Our emphasis on surprise and unpredictability—rather than minimizing error and surprise—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 significant value—needs to be unique and firm-specific (Felin and Zenger, 2016). And this firm-specificity is tied to unique beliefs and the development of a theory-based logic for creating value unforeseen by others. This also enables economic actors to "hack" 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).

Now, in emphasizing the role of human cognition when it comes to strategic decision making—and the associated development of theory-based causal logic—we certainly do not mean to minimize the role that AI can play in augmenting human decision making in various decision-related contexts. *An overwhelming number of decisions made by humans are quite routine and thus amenable to algorithmic processing* (Amaya and Holweg, 2024; Holmström, Holweg, Lawson, Pil, and Wagner, 2019). AI will undoubtedly play a major role in operations and various types of repeated decisions. Strategy decisions, on the other hand, are low-frequency and rare, while extremely impactful at the same time (Camuffo et al., 2022). But even in the context of these "exceptional" and rare-but-highly-impactful forms of decision making, we do see AI as serving a powerful function in assisting in the gathering, processing, and aggregation of information. The central argument we have laid out in this paper is that both AI and humans have their respective strengths, which must be considered in the context of judgment and decision making. We acknowledge the power of AI in making predictions with extant data, as well as the power of human cognition to engage in theory-based causal reasoning, experimentation and problem solving. Others have also recently proposed various human-AI hybrid solutions to enable better strategic decision making (e.g., Bell et al., 2024; Choudhary, Marchetti, Shrestha, and Puranam, 2023; Gregory et al., 2021; Kemp, 2023; Kim et al., 2023; Raisch and Fomina, 2023). We agree with the broad message of this work, that powerful opportunities lie in hybrid human-AI solutions that capitalize on the respective capabilities of each.

That said, 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. Our argument would be that a decision maker's (like a firm's) own theory needs to drive this process. AI that uses "universally"-trained AIs—based on widely-available models and training datasets—will yield general and non-specific outputs. There is a risk that off-the-shelf AI solutions will be susceptible to the general "productivity paradox" of information technology (Brynjolffson and Hitt, 1998), where investments in AI actually do *not* yield any productivity gains to those buying these tools. For AI to actually be a useful decision making tool, AIs can (and should) be customized, purpose-trained, and fine-tuned—they need to be made *specific*—to the theories, datasets and proprietary documents of decision makers like firms—thus yielding

unique outputs that reflect unique (rather than ubiquitously available) training data. It is here that we see significant opportunity for AI to serve as a powerful aid in strategic decision making.

By way of a final caveat, in this paper we have offered a theoretical argument for why and how AI differs from human cognition. That said, we recognize that AI is a fast-moving target, and thus our argument risks being outdated by rapidly emerging technologies and new understandings. However, our goal has been to demystify ongoing developments within AI and to provide some perspective on the respective strengths (and limitations) possessed by machines and humans. The excitement about AI is warranted, and there is no question that current and ongoing developments will fundamentally disrupt the nature of work and decision making in many domains (Amaya and Holweg, 2024). However, while the capabilities of AI (like LLMs) are remarkable, they presently lack the ability to engage in the type of forward-looking theorizing and the development of causal logic that is natural to humans. Furthermore, it is worth remembering that the current and ongoing accomplishments in AI are very much *human* accomplishments. The algorithms created to enable machine learning and prediction were generated by humans. The vast corpora of text and images used to train machines were created by humans. And humans of course are also playing a central role in the ongoing improvements to these models. While there are fears of an AI takeover, we think the "division of labor" between what AI does and what humans do will morph and evolve, as has been exemplified by the evolution of past technologies.

CONCLUSION

AI is anchored on data-driven prediction. We argue that AI's data and prediction-orientation is an incomplete view of human cognition. While we grant that there are some parallels between AI and human cognition—as a (broad) form of information processing—we focus on key differences. We specifically emphasize the forward-looking nature of human cognition and how theory-based causal logic enables humans to intervene in the world, to engage in directed experimentation, and to problem solve. Heterogeneous beliefs and theories enable the generation of *new* data (for example, through experimentation), rather than merely being reliant on prediction based on past data. That said, our argument by no means negates or questions many of the exciting developments within the domain of AI. We anticipate that AI will enable humans to make better decisions across many domains, particularly in settings that are conducive to routine, repetition, and computation. However, strategic decisions—given the emphasis on unpredictability and *the new*—provide a realm that is not readily amenable to data-based

prediction. Thus we question the idea that AI will replace human decision makers (e.g., Kahneman, 2018), and emphasize the importance of asymmetric beliefs and theory-based causal logic in human cognition.

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.

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.

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.

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.

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. 2023. Black boxes or unflattering mirrors? comparative bias in the science of machine behavior. *British Journal for the Philosophy of Science*, 74(3), 681-712.

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.

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.

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*.

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.

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., & 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.

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

Felin, T., Gambardella, A., Novelli, E., & Zenger, T. 2023. 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. 2023. Disruptive evolution: Harnessing functional excess, experimentation, and science as tool. *Industrial and Corporate Change*.

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.

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.

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

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.

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.

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*.

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.

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*.

Koenderink, J. 2012. Geometry of imaginary spaces. *Journal of Physiology*, *106*(5-6), 173-182.

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

Koenderink, J. J. 2012. Geometry of imaginary spaces. *Journal of Physiology*, 106(5-6), 173-182.

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.

Kramár, J., Eccles, T., Gemp, I., Tacchetti, A., McKee, K. R., Malinowski, M., & Bachrach, Y. 2022. Negotiation and honesty in artificial intelligence methods for the board game of Diplomacy. *Nature Communications*, 13(1), 7214.

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.

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., 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.

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.

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.

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.

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.

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

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. 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.

Uexküll, J.V., 1926. *Theoretical biology*, trans. by D.L. MacKinnon. New York: Harcourt, Brace and Company.

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.

Zaadnoordijk, L., Besold, T. R., & Cusack, R. 2022. Lessons from infant learning for unsupervised machine learning. *Nature Machine Intelligence,* 4(6), 510-520.

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

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., & Wen, J. R. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

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*.