Strategy Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Theory Is All You Need: AI, Human Cognition, and Causal Reasoning

Teppo Felin, Matthias Holweg

To cite this article:
Teppo Felin, Matthias Holweg (2024) Theory Is All You Need: AI, Human Cognition, and Causal Reasoning. Strategy Science 9(4):346-371. https://doi.org/10.1287/stsc.2024.0189

https://pubsonline.informs.org/journal/stsc

STRATEGY SCIENCE
Vol. 9, No. 4, December 2024, pp. 346–371
ISSN 2333-2050 (print), ISSN 2333-2077 (online)

Theory Is All You Need: AI, Human Cognition, and Causal Reasoning

Teppo Felin,[a,b,*] **Matthias Holweg**[b]

[a] Huntsman School of Business, Utah State University, Logan, Utah 84322; [b] Saïd Business School, University of Oxford, Oxford OX1 1HP, United Kingdom

*Corresponding author

Contact: teppo.felin@usu.edu, https://orcid.org/0000-0003-2044-0145 (TF); matthias.holweg@sbs.ox.ac.uk, https://orcid.org/0000-0001-9403-1681 (MH)

Received: February 23, 2024
Revised: April 20, 2024; July 3, 2024
Accepted: September 4, 2024
Published Online in Articles in Advance: December 3, 2024

https://doi.org/10.1287/stsc.2024.0189

Abstract. Scholars argue that artificial intelligence (AI) can generate genuine novelty and new knowledge and, in turn, that AI and computational models of cognition will replace human decision making under uncertainty. We disagree. We argue that AI's data-based prediction is different from human theory-based causal logic and reasoning. We highlight problems with the decades-old analogy between computers and minds as input–output devices, using large language models as an example. Human cognition is better conceptualized as a form of theory-based causal reasoning rather than AI's emphasis on information processing and data-based prediction. AI uses a probability-based approach to knowledge and is largely backward looking and imitative, whereas human cognition is forward-looking and capable of generating genuine novelty. We introduce the idea of data–belief asymmetries to highlight the difference between AI and human cognition, using the example of heavier-than-air flight to illustrate our arguments. Theory-based causal reasoning provides a cognitive mechanism for humans to intervene in the world and to engage in directed experimentation to generate new data. Throughout the article, we discuss the implications of our argument for understanding the origins of novelty, new knowledge, and decision making under uncertainty.

Keywords: cognition • artificial intelligence • prediction • causal reasoning • decision making • strategy • theory-based view

Introduction

Artificial intelligence (AI) now matches or outperforms humans in any number of games, standardized tests, and cognitive tasks that involve high-level thinking and strategic reasoning. For example, AI engines can readily beat humans in chess, which, for decades, served as a key benchmark of AI capability (Simon 1985a, Bory 2019). AI systems also now perform extremely well in complex board games that involve sophisticated negotiation, complex interaction with others, alliances, deception, and understanding other players' intentions (e.g., Ananthaswamy 2022). Current AI models also outperform more than 90% of humans in various professional qualification exams, such as the bar exam in law and the certified public accountant exam in accounting (Achiam et al. 2023). AI has also made radical strides in medical diagnoses, beating highly trained medical professionals in diagnosing some illnesses (e.g., Zhou et al. 2023). These rapid advances have led some AI scholar to argue that even the most human of traits, such as consciousness, will soon be replicable by machines (e.g., Goyal and Bengio 2022, Butlin et al. 2023). In all, AI is rapidly devising algorithms that "think humanly," "think rationally," "act humanly," and "act rationally" (Csaszar and Steinberger 2022, pp. 2–3).

Given the astonishing progress of AI, Daniel Kahneman (2018, pp. 609–610, emphasis added) 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."

Kahneman is not alone in this assessment. Davenport and Kirby (2016, p. 29) argue that "we already know that analytics and algorithms are better at creating insights from data than most humans" and "this human/machine performance gap will only increase." Many scholars claim that AI is likely to outperform

humans in most—if not all—forms of reasoning and decision making (e.g., Legg and Hutter 2007, Morris et al. 2023, Grace et al. 2024). Some argue that strategic decision making might also be taken over by AI (Csaszar et al. 2024) or even that science itself will be automated by "AI scientists" (e.g., Lu et al. 2024, Manning et al. 2024). One of the pioneers of AI, Geoffrey Hinton, argues that large language models (LLMs) already are sentient and intelligent and "digital intelligence" will inevitably surpass human "biological intelligence"—if it has not already done so (see Hinton 2023; also see Bengio et al. 2023).

Compared with machines, the cognitive and computational limitations of humans are obvious. Humans are biased (Kahneman 2011, Chater et al. 2018). Humans are selective about what data they attend to and sample, and they are susceptible to confirmation bias, motivated reasoning, and hundreds of other cognitive biases (nearly 200 as of last count). In short, humans are boundedly rational—significantly hampered by their ability to compute and process information (Simon 1955), particularly compared with computers (cf. Simon 1990). And the very things that make humans boundedly rational and poor at decision making are seemingly the very things that enable computers to perform well on cognitive tasks. The advantage of computers and AI is that they can handle vast amounts of data and process it quickly and in powerful ways.

In this paper, we offer a contrarian view of AI relative to human cognition, including its implications for strategy, the emergence of novelty, and decision making under uncertainty. AI builds on the idea that cognition—by both machines and humans—is a generalized form of information processing: a type of input–output device. To illustrate cognitive differences between humans and computers, we use the example of large language models versus human language learning. We introduce the notion of data–belief (a)symmetry and the role this, respectively, plays in explaining AI and human cognition, using heavier-than-air flight as an extended example. Human cognition is forward-looking, necessitating data–belief asymmetries, which are manifest in theories, causal reasoning, and experimentation. We argue that human cognition is driven by forward-looking, theory-based causal logic, which is distinct from the emphasis AI and computational models of cognition place on prediction and backward-looking data. Theory-based causal reasoning enables the generation of new and contrarian data, observations, and experimentation. We highlight the implications of these arguments for understanding the origins of novelty, new knowledge, and decision making under uncertainty.[1]

AI = Mind: Is Cognition Computation?

Modeling the human mind—thinking, rationality, and cognition—has been the central aspiration and ambition behind AI from the 1940s to the present (McCulloch and Pitts 1943, Turing 1948/1992; 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, p. 12). The commonalities between models of AI and human cognition are not just historical, but these linkages have only deepened in the intervening decades (for a review, see Sun 2023; also see Laird et al. 2017). Computation also underlies many other models of cognition, including the concept of mental models (Johnson-Laird 1983), the Bayesian brain, and predictive coding or processing (e.g., Friston and Kiebel 2009; Hohwy 2013, 2020). In fact, cognitive scientist Johnson-Laird (1983, p. 477) goes so far as to argue that "any scientific theory of the mind has to treat it as an automaton."

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

Now, our overall purpose is not to exhaustively review models of AI and cognition, particularly as excellent reviews can be found elsewhere (e.g., Goodfellow et al. 2016, Aggarwal 2018, Russell and Norvig 2022). Rather, we simply want to point out the strong emphasis that past, current, and ongoing research places on the similarities between models of AI and human cognition. To assure the reader that we are not creating a caricature of existing work, we have provided relevant, additional detail about AI-cognition similarities in the appendix.

There, we more exhaustively point out examples of how scholars—from the 1950s to the present—have sought to create an equivalence between AI, machines, and human cognition. In all of this work, cognition and computation (and AI) are seen as deeply connected: the underlying premise of this work is that machines and humans are a form of input–output device, in which the same underlying mechanisms of information processing and learning are at play. The focus on computation and information processing also is the axiomatic basis for the concept of bounded rationality (for a review, see Felin et al. 2017). Bounded rationality focuses on human "computational capacities" and their limits (Simon 1955, p. 99), and this idea has deeply shaped fields such as economics, decision theory, strategy, and the cognitive sciences (e.g., Chater et al. 2018, Kahneman 2003, Puranam et al. 2015, Gigerenzer and Goldstein 2024).

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

Machine vs. Human Learning: Different Inputs, Different Outputs

While the input–output model of minds and machines—whether we are talking about symbolic or subsymbolic approaches (see the appendix for further detail)—has been a central emphasis of AI and cognitive science, next we highlight some important *differences* between machine learning and human learning. An apt context for highlighting these differences is to focus on language. Language arguably is "the most defining trait of human cognition (language and its relation with thought)," and therefore, it "can be a true 'window into the mind'" (Chomsky and Gallega 2020, p. 321; also see Pinker 1994).[2] Language provides an important test and context for understanding human and artificial intelligence. Furthermore, some have already argued that large language models are sentient with a few even arguing that they already closely mirror or exceed human cognition (e.g., Binz and Schulz 2023, Hinton 2023)—an assumption that we challenge.

At the most basic level, to study any system and its behavior, we need to understand its inputs and outputs. Turing (1948/1992) 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 (1950, p. 456, emphasis added) made the analogy to an "untrained infant brain," saying an infant brain is largely a blank slate, "something like a notebook" with "*little* mechanism, and lots of *blank* sheets" (cf. Turing 1948/1992). According to Turing, these blank sheets are (or need to be) filled with inputs via the process of training and education. Through the early course of its life, an infant or child is taught and receives inputs in the form of language and spoken words that it hears and encounters. Education and training represent the inputs that eventually account for human linguistic capacities and outputs. And in the same way, Turing (1948/1992, p. 107) argues, one can think of an "analogous teaching process applied to machines," where machines learn from their inputs. Turing lists various settings in which a thinking machine might show that it has learned—including games such as chess or poker, cryptography, or mathematics—and he argues that the "learning of languages would be the most impressive, since it is the most human of these activities" (Turing 1948/1992, p. 117). As human and machine learning are often seen as a similar process, we next focus on key differences using language learning as our example. We then highlight the implications of these differences in learning for decision making and knowledge generation both in scientific and economic contexts.

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

The vast corpus of text used to train an LLM is tokenized to enable natural language processing. This typically involves converting words (or subword units or characters) into numerical sequences or vectors. To illustrate, a sentence such as "The cat sat on the mat" might be tokenized into a sequence such as [10, 123, 56, 21, 90, 78]. Each token is passed through an embedding layer, which converts the token into a dense vector

representation that captures semantic information, such as its frequency and positional embedding. The embedding layer has its own set of parameters (weights) that are learned during training. The attention mechanism introduced with the transformer architecture (Vaswani et al. 2017), touched on by us previously, allows the model to consider each token in the context of all other surrounding tokens and thus to gain an understanding of the wider context. Deep artificial neural networks have turned out to be extremely general and applicable not just to text, but to varied domains such as image recognition and computer vision, including multimodal applications that combine various types of data (for example, enabling the creation of images from text prompts).[3]

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

Now, whereas 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 Naveed et al. 2023, Chang et al. 2024, Minaee et al. 2024; also see Resnik 2024). The end result of this training is an AI model that is capable of language: more specifically, the model is capable of generating fluent and coherent text by using a stochastic approach of next word prediction in response to a prompt. In short, LLM outputs are based on conditional probabilities given the structure of the inputs they have encountered in their training data.

Based on this broad outline of how an LLM is trained, we compare this to how humans learn language. We should reiterate, as discussed at the outset of this article, that the basic premise behind models of AI is that there is a symmetry between how machines and humans learn. We think it is important to carefully point out differences as these provide the foundation for our subsequent arguments about cognition and the emergence of novelty.

How Humans Learn Language Compared with Machines. The differences between human and machine learning—when it comes to language (as well as other domains)—are stark. Whereas 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., Hart and Risley 2003, Gilkerson et al. 2017). 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) and also in terms of their quality.[5] Namely, the spoken language to which an infant or young child is (largely) exposed is different from the written language on which an LLM is trained. Spoken language differs significantly from written language in terms of its nature, structure, and purpose. Here, the research on the differences between spoken and written language is highly instructive (e.g., Biber 1991). Spoken language is spontaneous (not meaningfully edited), informal, repetitive, and often ephemeral. Written language—on the other hand—is visual and permanent, more carefully crafted, planned, and edited. It is also denser, featuring more complex vocabulary (e.g., Halliday 1989, Tannen 2007). Importantly, the functional purposes and uses of spoken versus written language also differ significantly. Spoken language is immediate, interactive, focused on coordinating, expressing, and practically doing things. Whereas written language also serves these purposes, the emphasis is more on the communication of complex information. The vast bulk of the training data of the LLM is not conversational (for models trained on spoken language or raw audio, see Lakhotia et al. 2021). Rather, written language is more carefully thought out. An LLM is likely to be trained with the works of Shakespeare and Plato, academic publications, public domain books (e.g., from Project Gutenberg), lyrics, blog posts, news articles, and various material from the internet. These data are far cleaner, far more correct grammatically, and organized. Arguably the inputs received by an LLM—in the form of written, edited, and published text—are linguistically far superior. In a statistical sense, LLM training data contain less noise and thus offer greater predictive power. Even the vast stores of Wikipedia articles that are included in most LLM training data sets are the end result of thousands of edits to ensure readability, accuracy, and flow.

Clearly, humans learn language under different conditions and via different types of inputs. In short, it can readily be argued that the human capacity for language develops differently from how machines learn language in both quantity and quality. Humans (somehow) learn language from extremely sparse, impoverished, and highly unsystematic inputs and data (Chomsky 1975). Compared with LLMs, human linguistic capabilities are radically "underdetermined" by the inputs. That is, the relatively sparse linguistic inputs can scarcely account for the radically novel outputs generated by humans.[6]

Beyond the quantitative and qualitative differences in inputs (when it comes language learning by LLMs

versus humans), it is important to compare the linguistic outputs and capabilities of machines versus humans. In terms of output, LLMs are said to be generative (the acronym GPT stands for generative pretrained transformer).

But in what sense are LLMs generative? They are generative in the specific sense that they are able to create novel outputs by probabilistically sampling from the vast combinatorial possibilities in the associational and correlational network of word frequencies, positional encodings, and co-occurrences encountered in the training data (Vaswani et al. 2017).[7] The LLM is generative in the sense that the text that is produced is not simply plagiarized or copied verbatim from existing sources contained in the pretraining data (McCoy et al. 2023). In the process of generating text, the parameters (weights and biases) determine how much influence different parts of the training data probabilistically have on the output. For example, in a sentence-completion task, the weights—developed from the corpus of the training data—help the model decide which words are most likely to come next based on the context provided by the input. The output is statistically derived (or, put differently, probabilistically drawn) from the training data's underlying linguistic structure. The outputs, therefore, have compositional novelty (in terms of novel ways of saying the same thing—more on this below), and they also manifest some analogical generalization (McCoy et al. 2023). That said, any assessment of how good an LLM is needs to recognize "the problem that [LLMs] were trained to solve: next-word prediction" (McCoy et al. 2024, p. 4). And as next-word prediction engines, LLMs certainly demonstrate exceptional capabilities.

Beyond Mirroring: Can AI Generate Genuine Novelty?

So far we have summarized the central elements of a particular AI system—an LLM—and compared it with humans. Next, we further address whether an AI can be said to be intelligent and whether it can generate genuine novelty and new knowledge. Whereas our focus remains on LLMs, we extend our arguments to other forms of AI and cognitive approaches that focus on data and prediction. We concurrently raise questions about whether an AI system meaningfully can originate new knowledge and engage in decision making under uncertainty.

AI: Intelligence and New Knowledge?

As we foreshadow above, an AI such as an LLM seems to mirror the inputs with which it has been trained rather than meaningfully manifest some form of intelligence. But beyond next-word prediction and linguistic fluency, could an LLM do a better job than humans in decision making under uncertainty (e.g., Csaszar et al. 2024; cf. Kahneman 2018), or could an LLM or AI scientist perhaps even automate science itself (e.g., Lu et al. 2024, Manning et al. 2024; also see Kıcıman et al. 2023; Agrawal et al. 2023, 2024)?[8]

Without question, LLMs seem to manifest sparks of intelligence. But intelligence is not simply memorization or the ability to restate or paraphrase information in various ways. We argue that LLMs appear intelligent because they capitalize on the fact that the same thing can be stated, said, and represented in indefinite ways. This is readily illustrated by the fact that the revolutionary breakthrough that gave rise to LLMs—the transformer architecture—was developed in the context of language translation (Vaswani et al. 2017). In an important sense, LLMs can be seen as translation generalized. They represent 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 with a different set of words or more abstractly: to represent the same thing in a different format. LLMs serve this representational and mirroring function remarkably well. This representational and mirroring function from language to language is generalized to a process that takes one way of saying something and generates another way of saying the same thing. Stochastic next-word prediction using conditional probabilities—based on the weights and parameters derived from vast training data sets—allows for surprisingly rich combinatorial outputs. The learning of the LLM is embodied in the relationships found between words that are sampled to enable stochastic generativity, in which the outputs mirror past inputs. With vast data, an LLM is good at probabilistically and fluently predicting the next word. But, as we discuss, the fluency with which LLMs seem to predict and generate outputs dupes us into seeing them as intelligent, as if they are engaging in far more than mere mirroring or translation.

Before revisiting our question of whether an AI such as an LLM could actually originate novelty or engage in some form of forward-looking decision making, it is worth highlighting metaphorical similarities between AI and cognitive architectures based on prediction. For example, consider a cognitive approach such as predictive processing (Pezzulo et al. 2024), which shares broad similarities with active inference, the free energy principle, the Bayesian brain, and predictive coding. At a high level, both LLMs and predictive processing seek to engage in a similar process, namely, error minimization and iterative optimization, in which the systems are essentially navigating a high-dimensional space to find a state that minimizes both error and surprise. LLMs learn from the training data, and predictive processing learns from its environment (cf. Hohwy 2020). LLMs aim to reduce the difference between their probabilistic

predictions (the next word in a sentence) and the actual outcomes (the real next word), thereby improving their accuracy. Predictive processing, as a cognitive theory, posits that the brain continuously predicts sensory input and minimizes the error between its predictions and actual sensory input. The capability of each to predict—whether a word or a perception—is a function of past inputs. Large language models seek to predict the most likely next word based on training data, and active inference seeks to predict the most likely next percept or action. Both approaches are wildly conservative (tied to past data) as they seek to reduce surprise—or to engage in prediction as error minimization (Hohwy 2013).[9] Back-propagation, a fundamental mechanism in training neural networks, and the concept of error minimization in predictive processing (Friston and Kiebel 2009) share a broad conceptual similarity in that both involve iterative adjustments to minimize some form of error or discrepancy. Both generate a prediction based on past inputs. Both back-propagation and error minimization in predictive processing involve adjusting an internal model (neural network weights in AI and hierarchical brain models in neuroscience) to reduce error (or, in machine learning terms, minimize the loss function).

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

We do not believe LLMs or input–output–based cognitive systems can do this—at least not beyond random flukes that might emerge because of their stochastic nature.[10] There is no forward-looking mechanism or unique causal logic built into these systems. It is important to clearly delineate why this is the case as some argue and anticipate that LLMs will replace human decision makers in uncertain contexts such as strategy and even science itself. For example, Csaszar et al. (2024) argue that "the corpora used to train LLMs include valuable information for SDM, such as consumer preferences, competitor information, and strategy knowledge" and point to how an AI can use various decision-making tools to generate business plans and strategy (Csaszar et al. 2024, p. 2). And Manning et al. (2024) even argue that LLMs will automate social science given their seeming ability to generate hypotheses and causal models, including testing them (also see Lu et al. 2024).

These claims are vastly overstated. One way to think about this is that a prediction-oriented AI such as an LLM can essentially be seen as possessing Wikipedia-level knowledge. On any number of topics (if contained in the training data), an LLM can summarize, represent, and mirror the words it has encountered in various different and new ways. On any given topic—again, if sufficiently represented in the training data—an LLM can generate indefinite numbers of coherent, fluent, and well-written Wikipedia articles by drawing on the conditional probabilities it has learned. But, just as a subject-matter expert is unlikely to learn anything new about the expert's specialty from a Wikipedia article within the expert's domain of expertise, an LLM is unlikely to somehow bootstrap knowledge beyond the combinatorial possibilities of the word associations it has encountered in the past. It has no forward-looking mechanism for doing so.

There is also good evidence to suggest that when an LLM encounters (is prompted with) a reasoning task, it merely reproduces the linguistic answers (about reasoning) it has encountered in the training data rather than engaging in any form of actual, on-the-fly reasoning. If the wording of a reasoning task—such as the Wason selection task or the Monty Hall problem—is changed only slightly, LLM performance declines significantly below human performance, and the mistakes of the LLM are glaringly obvious to humans (e.g., Hong et al. 2024). LLMs are not meaningfully engaged in any form of real-time reasoning (as assumed by Lu et al. 2024, Manning et al. 2024). Rather, they are merely repeating the word structures associated with reasoning, which they have encountered in the training data. This effect can also be shown empirically as training LLMs on their past output leads to a rapid decline in performance and even their collapse (Shumailov et al. 2024). Importantly, LLMs memorize and regurgitate the words associated with reasoning but do not engage in on-the-fly reasoning of any sort.[11] This is why Francois Chollet (2019) created the "abstraction and reasoning corpus" as a challenge or test to see if an AI system can actually solve new problems (that is, problems it has not encountered in its training data) without merely resorting to memorized answers and solutions encountered in the past (which captures the present state of AI systems, including LLMs).[12]

That said, our goal is not to dismiss the remarkable feats of LLMs or other forms of AI or applications of machine learning. The fact that an LLM can outperform most humans in varied types of tests and exams is remarkable (Achiam et al. 2023). But this is because it has encountered this information, memorized it, and is able to repeat it in fluent ways. An LLM essentially has a superhuman capacity for memorization and an ability to summarize memorized word structures in diverse ways. In all, certainly the idea of LLMs as "stochastic parrots" or "glorified auto-complete" (Bender et al. 2021) underestimates their ability. But, equally, ascribing LLMs the ability to actually reason and generate new knowledge vastly overestimates their ability. LLMs are essentially powerful and creative imitation engines in stochastically and probabilistically assembling words though not linguistically innovative compared with children (see Yiu et al. 2023). The idea that LLMs

somehow generate new-to-the-world knowledge—or feature something like human consciousness—seems to be a significant stretch (though, see Butlin et al. 2023, Hinton 2023). In sum, the generativity of these models is a type of lowercase "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 something novel—such as new knowledge—with an LLM, imagine the following thought experiment. Imagine an LLM in the year 1633, where the LLM's training data incorporates 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? Because the LLM would probabilistically sample from the association and correlation-based word structure of its vast training data—again, everything that has so far been written (including all the scientific writings about the structure of the cosmos)—it would only restate, represent, and mirror the accumulated scientific consensus. The training data set for the LLM would overwhelmingly feature texts with word structures supporting a geocentric view in the form of the work of Aristotle, Ptolemy, and many others. Ptolemy's careful trigonometric and geometric calculations, along with his astronomic observations, would be included in support of a geocentric view as represented in the many texts that would have summarized the geocentric view (such as de Sacrobosco's popular textbook *De saphera mundi*). These texts would feature word associations that highlight how the motions and movements of the planets could be predicted with remarkable accuracy with the predominant geocentric view. The evidence—as inferred from the repeated word associations found in the training data—would overwhelmingly be against Galileo. LLMs do not have any way of accessing truth (for example, through experimentation or counterfactuals) beyond mirroring and restating what is found in the text.

Even if alternative or heretical views were included in the training data (such as the work of Copernicus even though his work was largely banned), the logic of this work would be dwarfed by all the texts and materials that supported the predominant geocentric paradigm.[13] The overwhelming corpus of thousands of years of geocentric texts would vastly outweigh Galileo's view or anything supporting it. An LLM's model of truth or knowledge is solely statistical, relying on frequency and probability. Outputs are influenced by the frequency with which an idea is mentioned in the training data as reflected by associated word structures. For example, the frequency with which the geocentric view has been mentioned, summarized, and discussed in the training data necessarily imprints itself onto the output of the LLM as truth. As the LLM has no actual grounding in truth beyond the statistical relationships between words, it would say that Galileo's view and belief is delusional and in no way grounded in science.

A neural network such as an LLM might, in fact, include any number of delusional beliefs, including beliefs that turned out to eventually be correct (such as Galileo's) and also beliefs that objectively were (and still are) delusional. Ex ante, there is no way for an LLM to arbitrate between the two. For example, the eminent astronomer Tyco Brahe made and famously published extensive claims about astrology, the idea that celestial bodies and their movement directly impact individual human fates as well as political and other affairs. His astrological writings were popular not just among some scientists, but also among the educated elite. A hypothetical LLM (in 1633) would have no way of arbitrating between Galileo's (seeming) delusions about heliocentrism nor Brahe's (actual) delusions about astrology. Our hypothetical LLM would be far more likely to have claimed that Brahe's astrological claims are true than that Galileo's argument about heliocentrism is true. The LLM can only represent and mirror the predominant and existing conceptions—in this case, the support for the geocentric view of the universe—it finds in the frequencies and statistical association of words in its training data.

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

Some LLMs have sought to engineer around the problem of their frequency-based and probabilistic approach by creating so-called "mixture of experts" models in which the outputs are not simply the average result of outrageously large neural networks, but can be fine-tuned toward some forms of expertise (Shazeer et al. 2017, Du et al. 2022). Another approach is retrieval-augmented generation, which uses the general linguistic abilities of the LLM but limits the data used for prediction to a confined and preselected set of sources (Lewis et al. 2020). Furthermore, ensemble approaches—which combine or aggregate diverse architectures or outputs—have also been developed (Friedman and Popescu 2008, Russell and Norvig 2022). However, even here, the outputs would necessarily also be reflective of what any particular experts have said within the training data rather than any form of forward-looking projection or on-the-fly causal reasoning on the part of the LLM.

This problem is further compounded in situations that are characterized by high levels of uncertainty and novelty (such as many forms of decision making), in which the idea of expertise or even bounded rationality is hard to specify given an evolving and changing world (Felin et al. 2014).[14]

Finally, it is critically important to keep in mind that the inputs of any LLM are past human inputs, and therefore, outputs also roughly represent what we know so far. Inherently an LLM cannot go beyond the realms covered by the inputs. There is no mechanism to somehow bootstrap forward-looking beliefs about the future—or causal logic or knowledge—beyond what can be inferred from the existing statistical associations and correlations found in the words in the training data.

The Primacy of Data vs. Data–Belief Asymmetry

The central problem we have highlighted so far is that learning by machines and AI is necessarily backward looking and imitative. Again, this should not be read as a critique of these models, rather, merely as a description of their structural limits. Whereas they are useful for many things, an AI model—such as an LLM—is not able to generate new knowledge or solve new problems. An LLM does not reason. And an LLM has no way of postulating beyond what it has encountered in its training data. Next, we extend this problem to the more general emphasis on the primacy of data within both AI and cognitive science. Data itself, of course, is not the problem. Rather, the problem is that data are used in a theory-independent fashion (Anderson 2008). 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 in which data—such as cues, stimuli, text, images—essentially are read, learned, and represented by a system, whether it is a human or computational one. The world (any large corpus of images, text, or environment) has a particular statistical and physical structure, and the goal of a system is to accurately learn from it and reflect it. This is said to be the very basis of intelligence. As put by Poldrack (2021, p. 1307, emphasis added), "Any system that is going to behave intelligently in the world must contain *representations that reflect the structure of the world*" (cf. Yin 2020). Neural network–based approaches and machine learning with their emphasis on bottom-up representation offer the perfect mechanism for doing this because they can "learn directly from data" (Lansdell and Kording 2019; also see Baker et al. 2022). Learning is data-driven.[15] Of course, cognitive 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 et al. 2016).[16]

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

It is worth reflecting on the epistemic stance—or underlying theory of knowledge—that is presumed here. Knowledge is traditionally defined as justified belief, and belief is justified by data and evidence. As suggested by Bayesian models, we believe or know things to the extent to which we have data and evidence for them (Pinker 2021). Beliefs should be proportionate to the evidence at hand because agents are better off if they have an accurate representation or conception of their environment and the world (e.g., Schwöbel et al. 2018).[17] Knowledge can be seen as the accumulated inputs, data, and evidence that make up our beliefs. And the strength or degree of any belief should be symmetrical with the amount of supporting data or, put differently, the weight of the evidence (Pinker 2021; also see Griffin and Tversky 1992, Kvam and Pleskac 2016, Dasgupta et al. 2020). This is the foundation of probabilistic models of cognitive systems. These approaches focus on "reverse-engineering the mind"—from inputs to outputs—and they "[forge] strong connections with the latest ideas from computer science, machine learning, and statistics" (Griffiths et al. 2010, p. 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) in which humans or machines learn "through repeated interactions with an environment"—without "requiring any a priori specifications" (Binz et al. 2023). One way to summarize the above literature is that there needs to be a symmetry between one's belief and the corroborating

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

But what about edge cases? That is, what about situations in which an agent correctly takes in all the data and evidence yet somehow turns out to be wrong? Models based on rational information processing do not offer a mechanism for explaining change or new knowledge or an explanation of situations in which data and evidence-based reasoning might lead to poor outcomes (cf. Felin and Koenderink 2022). Furthermore, whereas learning-based models of knowledge enable belief updating based on new evidence, there is no mechanism for explaining where new data comes from or what data should be considered as relevant and what data should be ignored. And what if the data and evidence are contested? This is a particularly significant problem in contexts that feature rampant uncertainty, including any type of forward-looking decision making and scientific reasoning.

Explaining the emergence of novelty and new knowledge is highly problematic for computational, input–output models of cognition that assume what we call data–belief symmetry. The basis of knowledge is the quest for truth (Pinker 2021), which is focused on existing evidence and data. But we argue that data–belief *asymmetry*, in fact, is essential for the generation of new knowledge and associated decision making. The existing literature in the cognitive sciences focuses on one side of the data–belief asymmetry, namely, its downside: the negative aspects of data–belief asymmetries (e.g., Kunda 1990, Scheffer et al. 2022). This downside includes all the ways in which humans persist in believing something despite seemingly clear evidence to the contrary (Pinker 2021). This includes a large literature that focuses on human biases in information processing—the suboptimal and biased ways that humans process, perceive, and use data and fail to appropriately update their beliefs. This is evident in the vast literatures that focus on various data-related pathologies and biases, including motivated reasoning, confirmation bias, selective perception and sampling, and the availability bias. The emphasis on erroneous beliefs and human bias has powerfully influenced how we think about human nature and decision making within various social and economic domains (e.g., Kahneman 2011, Bénabou and Tirole 2016, Chater 2018, Gennaioli and Shleifer 2018, Kahneman et al. 2021, Bordalo et al. 2023).

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

To offer a practical and vivid illustration of how data–belief symmetry can be problematic, consider the beliefs that were held about the plausibility of heavier-than-air human-powered and controlled flight in the late 1800s and early 1900s. (We introduce this example here and revisit it throughout the remainder of the manuscript.) To form a belief about the possibility of human-powered flight—or even to 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 data point at the time was that human-powered flight was not a reality. This alone, of course, would not negate the possibility. So one might want to look at all the data related to human flight attempts to assess its plausibility. Here we would find that humans have tried to build flying machines for centuries, and flight-related trials had, in fact, radically accelerated during the 19th century. All of these trials of flight could be seen as the data and evidence we should use to update our beliefs about the implausibility of flight. All of the evidence clearly suggested that a belief in human-powered flight was delusional. A delusion can readily be defined as having a belief contrary to evidence and reality (Pinker 2021, Scheffer et al. 2022): a belief that does not align with accepted facts. In fact, the *Diagnostic and Statistical Manual of Mental Disorders*, 4th and 5th editions—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—naïvely, it was thought—pointed to birds as evidence for the belief that humans might also fly. This was a common argument.[18] But the idea that bird flight somehow provided hope and evidence for the plausibility of human flight was seen as delusional by scientists and put to bed by the prominent scientist Joseph LeConte (1888, p. 69), who argued that flight was "impossible, *in spite of* the testimony of birds." 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." According to LeConte, weight was the critical determinant of flight. With his data, he clearly pointed out that no bird above the weight of 50 pounds is able to fly and thus concluded that humans cannot fly. After all, large birds such as ostriches and emus are flightless. And even the largest flying birds, he argued—such as turkeys and bustards—"rise with difficulty" and "are evidently near the limit" (LeConte 1888, pp. 69–76). Flight and weight are correlated. To this, Simon Newcomb (1901, p. 435)—one of the foremost astronomers and mathematicians of the time—added that "the most numerous fliers are little insects, and the rising series stops with the condor, which, though having much less weight than a man, is said to fly with difficulty when gorged with food."

The emphasis that LeConte placed on the weight of birds to disprove the possibility of human-powered flight highlights one of the problems with data and belief updating based on evidence. It is hard to know what data and evidence might be relevant for a given belief or hypothesis. The problem is—as succinctly put by Polanyi (1958, p. 31)—that "things are not labeled evidence in nature." Is the fact that small birds can fly and large birds cannot fly relevant to the question of whether humans can fly? What is the relevant data and evidence in this context? Did flight have something to do with weight or size or with other features such as wings? Did it have something to do with the flapping of wings (as Jacob Degen hypothesized)? Or did it have something to do with wing shape, wing size, or wing weight? Perhaps feathers are critical to flight. In short, it is hard to know what data might be relevant and useful.

Of course, not all our beliefs are fully justified in terms of direct empirical data that we ourselves have verified. We cannot—nor would we want to—directly verify all the data and observations that underlie our beliefs and knowledge. More often than not, for our evidence, we rightly rely on the expertise, beliefs, or scientific arguments of others, which serve as testimony for the beliefs that we hold (Coady 1992, Goldman 1999). The cognitive sciences have also begun to emphasize this point. Bayesian and other probabilistic models of cognition have introduced the idea of the reliability of the source when considering what data or evidence one should use to update beliefs and knowledge (e.g., Hahn et al. 2018, Merdes et al. 2021). This approach recognizes that not all data and evidence is equal. Who says what does matter. The source of evidence needs to be considered. For example, scientific expertise and consensus are critically important sources of beliefs and knowledge.

This is readily illustrated by heavier-than-air flight. So what might happen if we *weight* our beliefs about the plausibility of human flight by focusing on reliable, scientific sources and consensus? In most instances, this is a rational strategy. However, updating our belief on this basis when it comes to heavier-than-air flight during this time period would further reinforce the conclusion that human-powered flight was delusional and impossible. Again, scientists such as LeConte and Newcomb argued that flight was impossible by pointing to seemingly conclusive data and evidence. And, not only should we update our belief based on this evidence, but we should also further weight that evidence by the fact that it came from highly prominent scientists with seemingly relevant knowledge in this domain. LeConte, for example, became the eventual president of the leading scientific association in the United States (the American Association for the Advancement of Science). And LeConte was scarcely alone. He was part of a much broader scientific consensus that insisted on the impossibility of human-powered flight. For example, Lord Kelvin emphatically argued—when 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 thermodynamics and hydrodynamics, the behavior of gases under different conditions (and other areas of physics), in fact, features practical implications that turned out to be extremely relevant for human-powered flight. And the aforementioned, prominent mathematician-astronomer Simon Newcomb (1901) also argued—in his article, "Is the airship coming?"—that the impossibility of flight was a scientific fact as there was no combination of physical materials that could be combined to enable human flight (for historical details, see Crouch 2002, Anderson 2004).

The question then is, how does someone still—despite seemingly clear evidence and scientific consensus—hold onto a belief that appears delusional? In the case of human flight, the data, evidence, and scientific consensus were firmly against the possibility. No rational Bayesian should have believed in heavier-than-air flight. Again, the evidence against it was not just empirical (in the form of LeConte's bird and other data) and based on science and scientific consensus (in the form of Kelvin and Newcomb's physics-related arguments), but it also was observationally salient. Many aviation pioneers not only failed and were injured, but some also died. For example, in 1896, the German aviation pioneer Otto Lilienthal died attempting to fly, a fact with which the Wright brothers were well acquainted (as they subsequently studied Lilienthal's notebooks and data). And, in 1903—just nine weeks before the Wright brothers succeeded—the scientist Samuel Langley failed spectacularly in his attempts at flight with large scientific and lay audiences witnessing the failures. Reflecting on recent flight attempts (including Langley's prominent failure), the editorial board of *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.

Now, we have, of course, opportunistically selected a historical example in which a seemingly delusional belief—one that went against existing data, evidence, and scientific consensus—turned out to be correct. Cognitive and social psychologists often engage in the "opposite" exercise in which they retrospectively point to situations in which humans doggedly persist in holding delusional beliefs despite clear evidence against those beliefs because of biased information processing, selective perception, or biased sampling of data (Festinger et al. 1956, Kunda 1990, Kahneman 2011, Pinker 2021; though see Anglin 2019). Conspiracy theories provide a frequently discussed example of beliefs that seem impervious to evidence (Gagliardi 2024, Rao and Greve 2024). Economists more generally have highlighted how humans can be "resistant to many forms of evidence, with individuals displaying non-Bayesian behaviors such as not wanting to know, wishful thinking, and reality denial" (Bénabou and Tirole 2016, p. 142).[19] Of course, some beliefs truly are delusional. But others—such as flight—may merely appear delusional.

We think that the other side of beliefs—beliefs that presently might appear delusional (beliefs that go against the evidence) and are seemingly driven by motivated reasoning but turn out to be correct—also need to be addressed. Our example of flight offers an instance of a far more generalizable process in which data–belief asymmetries are essential for the emergence of novelty and new knowledge. Heterogenous beliefs and data–belief asymmetries are the lifeblood of new ideas, new forms of experimentation, and new knowledge as we discuss next. Furthermore, this turns out to have important implications for computation-oriented forms of AI and cognition.

Theory-Based Causal Logic and Cognition

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

Our foundational starting point—building on Felin and Zenger (2017)—is that cognitive activity is a form of theoretical or scientific activity.[20] That is, humans generate forward-looking theories that guide their perception, search, and action. As noted by Peirce (1957, p. 71), 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." As highlighted by our example of language, the meager linguistic inputs of a child can scarcely account for the vast outputs, thus pointing to a human generative capacity to theorize. The human capacity to theorize—to engage in novel problem solving and experimentation—has evolutionary origins and provides a highly plausible explanation for evolutionary leaps and the emergence of technology (Felin and Kauffman 2023).

Importantly, theory-based cognition enables humans to do things, to experiment. This is also the basis of the so-called core knowledge argument in child development (e.g., Spelke et al. 1992, Carey and Spelke 1996). Humans develop knowledge as scientists do through a process of hypothesizing, causal reasoning, and experimentation. Whereas computational approaches to cognition focus on the primacy of data and environmental inputs, a theory-based view of cognition focuses on the active role of humans in not just learning about their surroundings, but also their role in actively generating new knowledge (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 practical or scientific knowledge. This is nicely captured in the title of an article in developmental psychology: "If You Want to Get Ahead, Get a Theory" (Karmiloff-Smith and Inhelder 1974). This also echoes Lewin's (1943, p. 118) maxim, "There is nothing as practical as a good theory." The central point here is that theories are not just for scientists. Theories are pragmatically useful for anyone seeking to understand and influence their surroundings; theories help us do things. Theorizing is a central aspect of human cognitive and practical activity. Thus, as argued by Dewey (1916, pp. 438–442), "the entities of science are not only from the scientist," and "individuals in every branch of human endeavor should be experimentalists." We build on this intuition and extend it into new and novel domains along with contrasting it with AI-informed models of cognition.

The theory-based view—in the context of decision making and strategy—extends the above logic and emphasizes the importance of theorizing and theories in economic contexts with widespread implications for cognition (Felin and Zenger 2017). The central idea behind the theory-based view is that economic actors

can (and need to) develop unique, firm-specific theories. Theories do not attempt to map existing realities, but rather to generate unseen future possibilities, and importantly, theories suggest causal interventions (experiments and actions that need to be taken) that enable the realization of these possibilities. Theories can also be seen as a mechanism for hacking competitive factor markets (cf. Barney 1986), enabling economic actors to see and search the world differently. Awareness for new possibilities is cognitively developed top-down (Felin and Koenderink 2022). Theories also have central implications for how to efficiently organize or govern the process of realizing something that is new (Wuebker et al. 2023). This approach has been empirically tested and validated (e.g., Camuffo et al. 2020, Novelli and Spina 2022, Agarwal et al. 2023), including important theoretical extensions (e.g., Ehrig and Schmidt 2022, Zellweger and Zenger 2023).[21] The practical implications of the theory-based view have also led to the development of managerial tools to assist start-ups, economic actors, and organizations in creating economic value (Felin et al. 2021a).

Our goal in this section of the paper is not to exhaustively review the theory-based view. Rather, our goal now is to further build out the cognitive and practical aspects of the theory-based view with a specific emphasis on causal reasoning and how this contrasts with backward-oriented, data-focused approaches to AI and cognition. We highlight how the human capacity for theorizing and causal reasoning differs from AI's emphasis on data-driven prediction. A theory-based view of cognition allows humans to intervene in the world beyond the given data—not just to process, represent, or extrapolate from existing data. Theories enable the identification or generation of nonobvious data and new knowledge through experimentation. This differs significantly from the arguments and prescriptions suggested by computational, Bayesian, 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 be superior to human judgment and cognition (e.g., Kahneman 2018).

Cognition: Data–Belief Asymmetry Revisited

Heterogeneous beliefs provide the initiating impetus for theory-based causal reasoning and cognition. From our perspective, for beliefs to be a relevant concept for understanding cognition and decision making, beliefs do not necessarily—in the first instance—need to be based on data. We are specifically interested in forward-looking beliefs, beliefs that presently lack evidence or even go against existing data but which might turn out to be true. Forward-looking beliefs, then, are more in search of data rather than based on existing data. At the forefront of knowledge, data are an outcome of beliefs—coupled with causal reasoning and experimentation (which we discuss in the next section)—rather than new knowledge being a direct outcome of existing data.

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

Before discussing how causal reasoning is essential for the realization of contrarian or delusional beliefs, it is worth emphasizing the 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 et al. 2021a). By way of contrast, the degree or strength of belief based on probabilistic or Bayesian models of cognition (cf. Pinker 2021) is directly tied to existing data and the weight of the available evidence (cf. Keynes 1921) rather than the likelihood of taking action—a significant difference.

Notice the implications of this in a context of our earlier example, human-powered flight. Belief played a central role in motivating action on the part of aviation pioneers despite overwhelming data and evidence against the belief. In a sense, those pursuing flight did not appropriately update their beliefs. Much, if not most, of the evidence was against the Wright brothers, but somehow, they still believed in the plausibility of flight. One of the Wright brothers, Wilbur, wrote to the scientist and aviation pioneer Samuel Langley in 1899 and admitted that "for some years I have been *afflicted with the belief* that flight is possible. My *disease* has increased in severity and I feel that it will soon cost me an increased amount of money if not my life" (Wright

and Wright 1881–1940, emphasis added). Wilbur clearly recognized that his belief about flight appeared delusional to others as is evident from his letters. But this belief motivated him to engage in causal reasoning and experimentation that enabled him and his brother to make the seemingly delusional belief a reality (only four short years later). Contrast the Wright brothers' belief with the belief of Lord Kelvin, one of the greatest scientific minds of the time. When invited to join the newly formed Aeronautical Society a decade earlier, Kelvin declined and said, "I have not the slightest molecule of faith in aerial navigation." Here Kelvin might have been channeling a scientific contemporary of his: the mathematician William Clifford (2010, p. 79), who argued that "it is wrong always, everywhere, and for anyone to believe anything on insufficient evidence." Kelvin did not want to lend support to what he considered an antiscientific endeavor. Without the slightest belief in the possibility of human flight, Kelvin naturally did not want to support anything that suggested human-powered flight might be possible. But, for the Wright brothers, the possibility of powered flight was very much a "live hypothesis" (James 1967). Despite the data, they believed human flight might be possible and took specific steps to realize their belief.

Asymmetries between data and beliefs present problems for the very idea of rationality (cf. Chater et al. 2018, Felin and Koenderink 2022). After all, to be a rational human being, our knowledge should be based on evidence. Our beliefs and knowledge should be proportionate to the evidence at hand. In a strict sense, the very concept of beliefs is not even needed as one can instead simply talk about knowledge, that is, beliefs justified by evidence. This is succinctly captured by Pinker (2021, p. 244), who argues, "I don't believe in anything you have to believe in." This seems like a reasonable stance. It is also the basis of Bayesian approaches in which 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 data–belief asymmetries—in which existing data presently does not corroborate beliefs or even goes against them—can be highly useful, even essential. They are the raw materials of technological and scientific progress. They are a central ingredient of decision making under uncertainty. Data–belief asymmetries direct our awareness toward new data and possible experiments to generate the evidence to support a belief. Of course, the idea of seeking data to verify a particular belief is the very definition of delusion and a host of associated biases, including confirmation bias, motivated reasoning, cherry-picking, denialism, self-deception, and belief perseverance. To an outsider, this looks like the perfect example of "the bad habit of seeking evidence that ratifies a belief and being incurious about evidence that might falsify it" (Pinker 2021, p. 13; also see Hahn and Harris 2014). Belief in human-powered flight readily illustrates this as there was plenty of evidence to falsify the Wright brothers' belief in the plausibility of heavier-than-air flight. Holding an asymmetric belief seems to amount to "wishful thinking" or "protecting one's beliefs when confronted with new evidence" (Kruglanski et al. 2020, p. 413; though see Anglin 2019). The Wright brothers were continuously confronted with evidence that disconfirmed their belief, including Samuel Langley's public failures with flight or the knowledge of Lilienthal's failed attempts (and his death because of a failed flight attempt). But, in these instances, ignoring the salient data and evidence—not updating beliefs based on seemingly strong evidence and even scientific consensus—turned out to be the correct course of action.

There are times when being (seemingly) irrational—ignoring evidence, disagreeing about its interpretation, or selectively looking for the right data—turns out to be the correct course of action. Human-powered flight, of course, is a particularly vivid illustration of this, though even more mundane forms of human behavior are fundamentally characterized by a similar process (Felin and Koenderink 2022). Most important for present purposes, our argument is that beliefs have a causal role of their own and can be measured by our propensity to act on them (Ramsey 1931, Felin et al. 2021a). Of course, having beliefs or having a willingness to act on them does not assure us that they are true. But they are an important motivation for action (Bratman 1987, Ajzen 1991).[22] And, again, notice that our emphasis on beliefs should not be seen as an attempt to dismiss the importance of data. Rather, as we highlight next, beliefs can motivate theory-based causal reasoning that directs human awareness toward actions and experiments that enable the generation of new data, evidence, and realization of new knowledge.

From Beliefs to Causal Reasoning and Experimentation

The realization of beliefs is not automatic. A central aspect of beliefs is their propensity to lead to causal reasoning and some form of directed experimentation. Beliefs enable actors to articulate a path for how to intervene in their surroundings and generate the evidence needed (Felin et al. 2021b). 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; also see Ehrig et al. 2024). This orientation toward intervention means that we do not simply take the world as it is; rather, we counterfactually think about possibilities and future states with an eye toward taking specific action, experimenting, and

generating the right evidence. This shifts the locus from backward-oriented information processing and prediction (in which the data are given) to doing and experimentation (in which the right data and evidence is identified or generated). This involves actively questioning and manipulating causal structures, allowing for a deeper exploration of what-if scenarios. Counterfactual thinking empowers humans to probe hypothetical alternatives and dissect causal mechanisms offering insights into necessary and sufficient conditions for an outcome (Felin et al. 2024). This approach is significantly different from input–output and information processing–oriented models of AI and computational cognition and various data-driven or Bayesian approaches to decision making. AI-based models of cognition largely focus on patterns based on past associations and correlations; prediction is based on past data. But these approaches lack an ability to understand underlying causal structures, hypothetical possibilities, and possible interventions (cf. Felin et al. 2021a, Ehrig et al. 2024). This is the role of theory-based causal logic.

A focus on plausible interventions and experimentation can be illustrated by extending our example of human-powered flight. This example also aptly illustrates the difference between how data-oriented and evidence-based scientists thought about the possibility of human-powered flight versus how more intervention-oriented and causal logic–based practitioners such as 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, and more importantly, they developed a causal theory of flight. Whereas 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.[23] And, whereas scientists had used bird data to argue that human flight was impossible (because of weight) (e.g., LeConte 1888, Newcomb 1901), the Wright brothers paid attention to a different aspect of bird flight. Ironically, bird-related data—though different aspects of it—provided seeming evidence for those advocating both for and against flight. LeConte focused on the weight of birds, whereas the Wright brothers engaged in observational studies of the mechanics of bird flight and anatomy (why birds were able to fly), for example, carefully studying the positioning of bird wings when banking and turning.

The key difference was that the Wright brothers with their belief in the plausibility of flight were building a causal theory of flying rather than looking for data that confirmed or disconfirmed whether flight was possible. The Wright brothers ignored the data and the scientific arguments of the naysayers. From the Smithsonian, the Wright brothers requested and received details about numerous historical flight attempts, including Otto Lilienthal's records. The Wright brothers notes and letters reveal that they carefully studied the flight attempts and aircraft of earlier pioneers such as George Cayley, Alphonse Penaud, and Octave Chanute (Wright and Wright 1841–1940, Anderson 2004, McCullough 2015). 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.

Again, the Wright brothers sought to develop their own, causal theory of flying. Their theory was not just motivated by their contrarian belief that flight was possible (a belief for which there did not seem to be any evidence). Their confidence in the plausibility of flight grew as they carefully studied the underlying mechanics of flight as they investigated the causal logic of flight. Most importantly, their causal reasoning led them to articulate the specific problems they needed to solve for human-powered flight to be possible. The Wright brothers reasoned that it was essential to solve three problems related to flight, namely, (a) lift, (b) propulsion, and (c) steering. To illustrate the power of developing a theory-based causal logic and identifying specific problems to solve, coupled with directed experimentation, we briefly discuss how they addressed one of the problems: the problem of lift.

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

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

The Wright brothers' approach to flight offers a useful case study and microcosm of how theory-based causal logic enables belief realization even when beliefs seemingly are not supported by existing data, evidence, or science. Based on their theorizing, study, and causal reasoning, the Wright brothers engaged in directed experimentation to solve the central problems of lift, propulsion, and steering. Their approach exemplifies the application of causal logic to understand and intervene in the world in the seeming absence of data (or even when data are contrary to one's belief). Their success with flight demonstrates how a systematic, intervention-oriented approach can unravel the causal mechanisms underlying complex phenomena and overcome the shortcomings of existing data.

As is implied by our arguments, we think scientific, economic, and technological domains are replete with opportunities for those with asymmetric beliefs to utilize theory-based causal reasoning and engage in directed experimentation and problem solving (Felin and Zenger 2017). As we argue, existing theories of cognition are overly focused on data–belief symmetry rather than data–belief asymmetry and how the latter enables causal reasoning that can enable the emergence of heterogeneity and the creation of novelty and value. Whereas there is much excitement about using AI to automate the generation of new knowledge and novelty generation (e.g., Csaszar et al. 2024, Lu et al. 2024, Manning et al. 2024) and even calls to replace biased human decision making by AI (e.g., Kahneman 2018), we argue that human causal reasoning cannot, at least presently, be mimicked by AI systems or computational approaches to cognition. Next, we further explore the implications of this argument for decision making under uncertainty and strategy.

Discussion: The Limits of Prediction for Decision Making Under Uncertainty

As we extensively discuss in this article, AI and the cognitive sciences use many of the same metaphors, tools, methods, and ways of reasoning about intelligence, rationality, and the mind. The prevailing assumption in much of the cognitive sciences is that the human mind is a computational input–output system (Christian and Griffiths 2016). Computational and algorithmic systems emphasize the power of prediction based on past data. The centrality of prediction is echoed by one the pioneers of AI, LeCun (2017), who argues that "prediction is the essence of intelligence."

Clearly, the predictive capabilities of AI are powerful. But is prediction central for decision making under uncertainty as well (that is, in unpredictable situations)? Many argue that this is the case (e.g., Davenport and Kirby 2016, Kahneman 2018). For example, in their book, *Prediction Machines*: *The Simple Economics of Artificial Intelligence*, Agrawal et al. (2022, pp. 22–32) emphasize that, stripped down to its essence, "AI is a *prediction* technology." And a central claim of their book is that "prediction is at the heart of making decisions *under uncertainty*" (Agrawal et al. 2022, p. 7, emphasis added). One way to summarize our argument in this paper is that we disagree with the importance placed on prediction—particularly in the form it is manifest in AI (that is, prediction based on past data)—especially in situations of uncertainty. Because the emphasis on prediction is commonplace, it is worth carefully pinpointing why we disagree with the importance placed on prediction.

Agrawal et al.'s (2022) argument offers a useful way for us to crystallize our more general concerns with the emphasis that is placed on prediction. Their argument might be summarized by pointing to a relatively common causal chain (of sorts), one that proceeds from data to information to prediction and to a decision or, in short, data → information → prediction → decision.[24] They specifically argue that "data provides the information that enables a prediction," and prediction, in turn, is "a key input into our decision making." This causal chain—from data to information to prediction and decision—certainly has intuitive appeal and mirrors what AI systems are good at: taking in vast amounts of inputs and data, processing this information, and then making predictions that can be used to make decisions. In short, as emphasized by Agrawal et al. (2022) and many others, data-driven prediction is at the heart of not just language models but AI more generally and also placed center stage in cognition.

But as we highlight throughout this paper, the problem is that data—data that is presently available or given—is not likely to be the best source of information and prediction when making forward-looking decisions. Data are snapshots of or mirrors to the past. Even vast amounts of data are unlikely to somehow enable one to anticipate the future (Felin et al. 2014). What is needed is some mechanism for projecting into the future and identifying the relevant data and evidence or, more likely, experimentally generating new data. This is the

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

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

Notice that our focus on unpredictability and surprise does not mean that we are somehow outside the realms of science or data. Quite the contrary. The process of making forward-looking decisions is about developing an underlying theory-based causal logic of how one might intervene in the world: essentially, outlining a causal path of how one might get from point A (the current state of the world) to point B (a hypothesized future state of the world). Theories create salience for the right interventions, experiments, and new data that enables the realization of beliefs that initially appear implausible. Theories play a central role in generating salience for what can be observed; the very idea of data (or observation) is theory dependent. 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, p. 604). Salience to the right (or new) data or forms of experimentation is given by a theory, not by past data. In this sense, theories can be said to have a predictive function, though here prediction is not a data-driven or error-minimizing process as it has been defined and operationalized within the context of AI (Agrawal et al. 2022) and cognitive science (cf. Clark 2018). Now, if the task at hand is routine and mundane—for example, "predict the next word in this sentence" or "tell me what you expect to see next"—then prediction with existing data can be useful. But the theory-based view is more focused on the forward-looking aspects of cognition and how human agents realize beliefs by developing a multistep causal path that enables the realization of beliefs through experimentation and problem solving. This is precisely what our example of the Wright brothers' theory of flying—and causal reasoning and experimentation—illustrates. It serves as microcosm of a far more general process of how humans intervene in their surroundings and realize novel beliefs. The economic domain is full of examples of how economic actors engage in this process (Felin and Zenger 2017).

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

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

that enables experimentation and the generation of new data and novelty.

Future Research Opportunities

The above arguments suggest a number of research opportunities, particularly when it comes to understanding AI, the emergence of novelty, and decision making under uncertainty. First, there is an opportunity to study when and how AI-related tools might be utilized by humans (such as economic actors) to create new value or to aid in decision making. If AI as a cognitive tool is to be a source of competitive advantage, it has to be utilized in unique or firm-specific ways. AI that uses universally available training data necessarily yields generic and nonspecific outputs. There is the risk that off-the-shelf AI solutions are susceptible to the "productivity paradox" of information technology (Brynjolfsson and Hitt 1998), in which investments in AI actually do not yield any gains to those buying these tools (rather only to those selling these technologies). Thus there is an opportunity to study how a specific decision maker's—such as a firm's—own theory of value can drive the process of AI development and adoption. For AI to actually be a useful tool for strategy and decision making, AI needs to be customized, purpose-trained, and fine-tuned—it needs to be made specific—to the theories, unique causal reasoning, data sets, and proprietary documents of decision makers such as firms. For example, advances in retrieval-augmented generation seem to offer a promising avenue to enhance specificity when using AI in strategic decision making. Any adoption of AI should be deliberate about which corpora and training data are utilized (and which not) when seeking unique AI-driven outputs. After all, the outputs of an AI—tailored to use specific data—are also the product of human agents who make decisions about which data are relevant and (which are not) for the decision at hand. It is here that we see an opportunity to understand how humans might uniquely interact with AI to generate these tools and associated human–AI interfaces. Early work has begun to look at how firms utilize AI to increase innovation or how various human–AI hybrid solutions enable better decision making (e.g., Gregory et al. 2021, Clough and Wu 2022, Choudhary et al. 2023, Girotra et al. 2023, Kemp 2023, Babina et al. 2024, Bell et al. 2024, Jia et al. 2024, Kim et al. 2024, Raisch and Fomina 2024, Tranchero et al. 2024). But there are promising opportunities to study how a particular economic actor's or firm's own theory and causal logic—as well as their unique or firm-specific sources of data and information—can shape the development or adoption of AI-related tools for executing strategy and making decisions.

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

Third, our arguments point to perhaps more foundational questions about the very nature of humans, particularly related to the purportedly computational nature of human cognition. Whereas questions about the nature of cognition might sound overly abstract and philosophical, they are critically important as they have downstream consequences for the assumptions we make and the methods we employ. Here we echo Simon (1985b, p. 303, emphasis added) who argued that "*nothing* is more important in setting our research agenda and informing our research methods than *our view of the nature of the human beings* whose behavior we are studying." So what is the predominant view of human cognition within AI and the cognitive sciences (and, by extension, in economics and strategy)? The predominant view of humans

is that they are input–output devices engaged in information processing akin to computers. In this paper we point out problems with the decades-old computer metaphor of the human mind, brain, and cognition. The computer has served as a central, organizing metaphor of human cognition for well over seven decades from the work of Alan Turing and Herbert Simon to modern instantiations of artificial neural networks, predictive processing, and the Bayesian brain (e.g., Knill and Pouget 2004, Cosmides and Tooby 2013, Kotseruba and Tsotsos 2020, Russell and Norvig 2022, Sun 2023, Gigerenzer and Goldstein 2024). A generalized computational approach to cognition, however, does not take into consideration the comparative nature of the organism under study because humans, organisms, and machines are all seen as the same—as invariant (see Simon 1990; cf. Simon 1980, Gershman et al. 2015). But there are significant differences in cognition, and these differences deserve careful attention. For example, computers do not meaningfully make decisions about which inputs might be relevant and which might not, nor can they meaningfully identify a new input, whereas humans have control over which inputs they might select or generate in the first place (Yin 2020, Brembs 2021, Felin and Koenderink 2022). Human cognition, as we discuss, is a form of forward-looking theorizing and causal reasoning. Notice that we are not trying to argue for some kind of human exceptionalism here, as these capacities are manifest—in different ways—across biological organisms more broadly (Riedl 1984; cf. Popper 1991).[25] There are significant research opportunities to study the endogenous and comparative factors that enable biological organisms and economic agents to theorize, reason, and experiment and to compare various forms of biological intelligence with artificial and nonbiological forms (cf. Levin 2024). Treating all cognition and intelligence as generalized computation unnecessarily narrows the scope of theoretical and empirical work and fundamentally misses the rich and heterogeneous ways that intelligence manifests itself across systems. Furthermore, the interfaces between biological and nonbiological forms of intelligence—as is manifest in the human use of technology and tools in evolution (Felin and Kauffman 2023)—provide intriguing opportunities for future work.

Conclusion

In this article we focus on the differences in cognition between AI and humans. Whereas AI-inspired models of cognition continue to emphasize the similarities between machines and humans, we argue that AI's emphasis on prediction (using past data) does not capture human cognition; that is, it cannot explain the emergence of novelty or new knowledge, nor can it assist in decision making under uncertainty. Overall, we grant that there are some parallels between AI and human cognition. But we specifically emphasize the forward-looking nature of human cognition and how theory-based causal reasoning allows humans to intervene in the world, to engage in directed experimentation, and to develop new knowledge. Heterogeneous beliefs and theories—data–belief asymmetries—enable the identification or generation of new data (for example, through experimentation) rather than merely being reliant on prediction based on the past data. AI-based computational models are necessarily built on data–belief symmetries. AI, therefore, cannot causally map and project into or anticipate the future as illustrated by LLMs. That said, our arguments by no means negate or question many of the exciting developments within the domain of AI. We anticipate that AI will help humans make better decisions across many domains, especially in settings that are characterized by routine and repetition. However, decisions under uncertainty—given the emphasis on unpredictability, surprise, and the new—provide a realm that is not readily amenable to data- or frequency-based prediction and associated computation. Thus we fundamentally question the notion that AI will (or should) replace human decision making (e.g., Kahneman 2018). We argue that humans—compared with computers and AI—have unique cognitive capacities that center on forward-looking beliefs and theorizing: the ability to engage in novel causal reasoning and experimentation.

Acknowledgments

Arguments related to this paper were presented at the *Strategy Science* "Theory-Based View" conference at Bocconi University as well as Harvard Business School, Aalto University, and the University of Illinois Urbana-Champaign. The authors are grateful for feedback from many participants and audience members that have helped improve their arguments. This paper is also much improved because of feedback from editors and peer review.

Appendix. AI and Human Cognition: Some Further Background

The earliest attempts to develop machines that simulate human thought processes and reasoning focused on general problem solving. Newell and Simon's (1963) general problem solver (GPS) represented an ambitious effort to (try to) solve any problem that could be presented in logical form. GPS used means–ends analysis, a technique that compared a current state to the desired state (or goal), identified the differences, and then applied operators (actions) to reduce these differences. The early excitement associated with GPS and other AI models—and their ability to mimic human intelligence and thought—was pervasive. As put by Herbert Simon in 1958, "There are now in the world machines that think, that learn and create. Moreover, their ability to do these things is going to increase rapidly until—in a visible future—the range of problems they can handle will be coextensive with the

range to which the human mind has been applied" (Simon and Newell 1958, p. 8).

Early models such as GPS provided the foundations for general cognitive architectures such as SOAR and ACT-R (Laird et al. 1987, Anderson 1990). The enthusiasm for these general models of cognition and AI continues to this day. Kotseruba and Tsotsos (2020) offer an extensive survey of more than 200 different cognitive architectures developed over the past decades. The ultimate goal of all this research into cognition, they argue, "is to model the human mind, eventually enabling us to build human-level artificial intelligence" (Kotseruba and Tsotsos 2020, p. 21). However, while various cognitive architectures related to AI hope to be general—and to mimic or even exceed human capability—their application domains have turned out to be extremely narrow and specific in terms of the problems they actually solve. But, despite limited success in generalizing early models of AI (specifically, from the late 1950s to the 1990s), excitement about the possibility of computationally modeling human cognition did not wane. Simon's frequent collaborator, Alan Newell (1990, p. 40), argued that "psychology has arrived at the possibility of unified theories of cognition," specifically in which "AI provides the theoretical infrastructure for the study of human cognition." This unified approach builds on the premise that humans share certain "important psychological invariants" with computers and artificial systems (Simon 1990, p. 3). This logic is also captured by such ideas as "computational rationality" (Gershman et al. 2015).

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

Most of the above attempts to model the human mind and mimic human reasoning focused on symbolic systems, so-called good old-fashioned AI. These approaches are an attempt to model thinking and intelligence through 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 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 prominent 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 such as this was to mimic the judgments of an expert decision maker. The model was a type of inference engine that used various preprogrammed 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 it emphasizes bottom-up learning. These models also see the mind (or brain) as an input–output device. But the emphasis is on learning things from scratch, that is, learning directly from data. Vast inputs and raw data are fed to these systems to recognize correlations and statistical associations or, in short, patterns. The weakness of the aforementioned symbolic systems is that these approaches are only useful for relatively static contexts that do not meaningfully allow for any form of dynamic, bottom-up learning from data or environments.

The foundations of subsymbolic AI were laid by scholars seeking to understand the human brain, particularly perception. Rosenblatt (1958, 1962; building on Hebb 1949) proposed one of the earliest forms of a neural network in his model of a "perceptron," which is the functional equivalent of an artificial neuron. Rosenblatt's work on the perceptron aimed to replicate the human neuron, which, when coupled together, would resemble human neural networks. Because modern artificial neural networks—including convolutional, recurrent autoencoders and generative adversarial networks—build on this broad foundation (e.g., LeCun et al. 2015, Aggarwal 2018), it is worth briefly highlighting the general architecture of this approach. The architecture of the multilayer 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 (Rumelhart et al. 1986, Hinton 1992, Bengio et al. 2021) and the basis for the radical advances made in areas such as AI image recognition and computer vision (Krizhevsky et al. 2012). While these models emerged seemingly out of nowhere, it is important to understand that the foundations were laid decades ago (Buckner 2023).

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 determines the perceptron's output based on the aggregated inputs from the association cells. The

perceptron's learning mechanism is based on feedback signals between the response units and the association units, allowing the network to learn and self-organize through repeated exposure to stimuli. So-called Hebbian learning (Hebb 1949), which posits the relatively cliché but important idea that "neurons that fire together, wire together," was the precursor to these types of feedback-based learning processes and many modern concepts of neural network theory.

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

In all, in this appendix we have sought to further highlight the deep connections between AI and computational models of human cognition. AI and other cognitive systems are treated in similar fashion as information processing machines or input–output devices (Simon 1990). While there has been widespread emphasis on the similarities between machines and humans, in this paper we explicitly focus on the differences and emphasize the importance of theory-based causal reasoning in human cognition.

Endnotes

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

[2] Recent comparisons between LLMs and humans reveal intriguing insights into formal versus functional linguistic competence. In humans, these two forms of competence rely on different neural mechanisms (Mahowald et al. 2024).

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

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

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

[6] This logic is aptly captured by Chomsky (1975, p. 179, emphasis added): "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." Our goal is not to endorse Chomsky's theory of universal grammar. Rather, we concur with this specific quote in terms of its characterization of the input–output relationship, in which human linguistic outputs are underdetermined by the inputs children receive. Broadly this also links to the alternative approach on which we focus (the theory-based view of cognition), discussed in the second half of the paper.

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

[8] AI can, of course, be (and has been) a powerful aid in scientific discovery. For example, modern AI techniques have analyzed astronomical data sets far more quickly and accurately than humans, helping identify new planets and celestial phenomena. Similarly, DeepMind's AlphaFold has revolutionized protein structure prediction, a critical task for understanding biological processes and

developing new medications (e.g., Jumper et al. 2021). Yet it is important to state that, in both of these cases, AI is not somehow independently doing the science by forming hypotheses or conducting experiments, but these hypotheses were provided by human scientists in the form of patterns and reward functions, respectively. AI has significantly accelerated research by enabling scientists to process large data sets and uncover novel patterns, allowing scientists to focus on hypothesis generation and experimental design.

[9] This leads to the problem of surprise and the famous dark room problem of predictive processing. For an attempt to deal with this, see Clark (2018; also see Constant et al. 2024).

[10] Though we, of course, recognize that there is significant disagreement on this point (for example, related to AI versus human creativity, see Franceschelli and Musolesi 2023).

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

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

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

[14] A deeper issue here is the frame problem (McCarthy and Hayes 1969) and the implications it has not just for artificial intelligence, but also for understanding decision making under uncertainty (Felin et al. 2014). In the latter context, the frame problem refers to the challenge of determining which aspects of a situation are relevant or irrelevant when making decisions (cf. Felin and Koenderink 2022). The frame problem highlights the difficulty of specifying all the possible consequences of an action in a dynamic environment, particularly when only some aspects of the world are affected by the action and others remain unchanged. In decision making under uncertainty, the frame problem underscores the complexity of reasoning about the implications of actions when the system must account for numerous potential variables and outcomes, often leading to difficulties in efficiently processing information and making reliable predictions.

[15] The problems with this approach are not just discussed by us. For example, see Yin (2020) for related points in the field of neuroscience.

[16] Machine-learning is said to be theory-free (to learn directly from data). However, 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 such as the choice of data, model architecture, and hyperparameter settings, as well as loss functions and metrics, regularization and generalization techniques, valued outputs, and types of human reinforcement.

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

[18] 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, p. 820).

[19] In economics, there is similarly an emphasis on how beliefs lead to negative outcomes. For example, Gennaioli and Shleifer's (2018; also see Bénabou and Tirole 2016) theory of beliefs focuses on beliefs that turn out to be delusional and are the result of poor judgment, biased information processing, and selective perception. In a related vein, Bordalo et al (2023) largely argue that humans are poor statisticians—selectively attending to and inappropriately weighting evidence and feedback—leading to suboptimal outcomes. In this paper we focus on discrepant or heterogeneous beliefs that appear delusional and highly biased to some, or even a majority, of actors in the present but turn out to be correct. Importantly, our theory is one of belief asymmetry rather than bounded rationality, bias, or information asymmetry (cf. Felin et al. 2024).

[20] A central aspect of this argument, which we unfortunately do not have room to explicate in this paper, is that humans are biological organisms. The theory-based view builds on the idea that all organisms engage in a form of forward-looking problem solving. A central aspect of this approach is captured by the biologist Rupert Riedl (1984, p. 8), who argued that "every conscious cognitive process will show itself to be steeped in theories; full of hypotheses." To see the implications of this biological argument on human cognition—particularly in comparison with statistical and computational approaches—see Felin and Koenderink (2022; also see Roli et al. 2022, Jaeger et al. 2024). For the embodied aspects of human cognition, see Mastrogiorgio et al. (2022).

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

[22] 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 in the context of humans, theory-based causal logic) as this has manifest in computational, Bayesian, and probabilistic forms of AI and cognition.

[23] 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, whereas 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. These data were critical in designing their successful aircraft.

[24] This has parallels with the data–information–knowledge–wisdom framework. For discussions of this see Felin et al. (2021b) and Yanai and Lercher (2020).

[25] For example, even simple organisms such as *Drosophila* (fruit flies) exhibit novel and surprising behaviors, such as initiating activity, expectations, and problem solving, that cannot be explained by or reduced to environmental inputs, genetic factors, or neural processing (see Heisenberg 2014).

References

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, McGrew B (2023) GPT-4 technical report. Preprint, submitted March 15, https://arxiv.org/abs/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. Preprint, submitted April 20, https://dx.doi.org/10.2139/ssrn.4412041.

Aggarwal CC (2018) *Neural Networks and Deep Learning* (Springer Publishing, New York).

Agrawal A, Gans J, Goldfarb A (2022) *Prediction Machines (Updated and Expanded): The Simple Economics of Artificial Intelligence* (Harvard Business Review Press, Boston).

Agrawal A, McHale J, Oettl A (2023) Superhuman science: How artificial intelligence may impact innovation. *J. Evolutionary Econom.* 33(5):1473–1517.

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

Ajzen I (1991) The theory of planned behavior. *Organ. Behav. Human Decision Processes* 50(2):179–211.

Amaya J, Holweg M (2024) Using algorithms to improve knowledge work. *J. Oper. Management* 70(3):482–513.

Ananthaswamy A (2022) DeepMind AI topples experts at complex game Stratego. *Nature* 604(7907):36.

Anderson JR (1990) *The Adaptive Character of Thought* (Psychology Press, London).

Anderson JD (2004) *Inventing Flight: The Wright Brothers and Their Predecessors* (Johns Hopkins University Press, Baltimore).

Anderson C (2008) The end of theory: The data deluge makes the scientific method obsolete. *Wired* 16(7).

Anglin SM (2019) Do beliefs yield to evidence? Examining belief perseverance vs. change in response to congruent empirical findings. *J. Experiment. Soc. Psych.* 82:176–199.

Babina T, Fedyk A, He A, Hodson J (2024) Artificial intelligence, firm growth, and product innovation. *J. Financial Econom.* 151:103745.

Baker B, Lansdell B, Kording KP (2022) Three aspects of representation in neuroscience. *Trends Cognitive Sci.* 26(11):942–958.

Barney JB (1986) Strategic factor markets: Expectations, luck, and business strategy. *Management Sci.* 32(10):1231–1241.

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

Bénabou R, Tirole J (2016) Mindful economics: The production, consumption, and value of beliefs. *J. Econom. Perspect.* 30(3):141–164.

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? *Proc. 2021 ACM Conf. Fairness Accountability Transparency* (ACM, New York), 610–623.

Bengio Y, Lecun Y, Hinton G (2021) Deep learning for AI. *Comm. ACM* 64(7):58–65.

Bengio Y, Hinton G, Yao A, Song D, Abbeel P, Harari YN, Hadfield G, Russell S, Kahneman D, Mindermann S (2023) Managing AI risks in an era of rapid progress. Preprint, submitted October 26, https://arxiv.org/abs/2310.17688.

Bhatia S (2023) Inductive reasoning in minds and machines. *Psych. Rev.* 130(1):105–125.

Biber D (1991) *Variation Across Speech and Writing* (Cambridge University Press, Cambridge, MA).

Binz M, Schulz E (2023) Turning large language models into cognitive models. Preprint, submitted June 6, https://arxiv.org/abs/2306.03917.

Binz M, Dasgupta I, Jagadish AK, Botvinick M, Wang JX, Schulz E (2023) Meta-learned models of cognition. *Behav. Brain Sci.* 47:e147.

Bordalo P, Conlon JJ, Gennaioli N, Kwon SY, Shleifer A (2023) How people use statistics. NBER Working Paper No. 31631, National Bureau of Economic Research, Cambridge, MA.

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

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

Brembs B (2021) The brain as a dynamically active organ. *Biochemical Biophysical Res. Comm.* 564:55–69.

Brynjolfsson E, Hitt LM (1998) Beyond the productivity paradox. *Comm. ACM* 41(8):49–55.

Buchanan BG, Shortliffe EH (1984) *Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project* (Addison-Wesley, Reading, MA).

Buckner CJ (2023) *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us About the Future of Artificial Intelligence* (Oxford University Press, Oxford, UK).

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. Preprint, submitted August 17, https://arxiv.org/abs/2308.08708.

Camuffo A, Gambardella A, Pignataro A (2022) Microfoundations of low-frequency high-impact decisions. Preprint, submitted June 23, https://dx.doi.org/10.2139/ssrn.4144724.

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

Carey S, Spelke E (1996) Science and core knowledge. *Philos. Sci.* 63(4):515–533.

Chang Y, Wang X, Wang J, Wu Y, Yang L, Zhu K, Xie X (2024) A survey on evaluation of large language models. *ACM Trans. Intelligent Systems Tech.* 15(3):1–45.

Chater N (2018) *Mind Is Flat: The Remarkable Shallowness of the Improvising Brain* (Yale University Press, New Haven, CT).

Chater N, Felin T, Funder DC, Gigerenzer G, Koenderink JJ, Krueger JI, Todd PM (2018) Mind, rationality, and cognition: An interdisciplinary debate. *Psychonomic Bull. Rev.* 25:793–826.

Chollet F (2019) On the measure of intelligence. Preprint, submitted November 5, https://arxiv.org/abs/1911.01547.

Chomsky N (1975) *Reflections on Language* (Pantheon, New York).

Chomsky N, Gallego ÁJ (2020) The faculty of language. *Revista Española de Lingüística* 50(1):7–34.

Choudhary V, Marchetti A, Shrestha YR, Puranam P (2023) Human-AI ensembles: When can they work? *J. Management*. 49(2):428–456.

Christian B, Griffiths T (2016) *Algorithms to Live By: The Computer Science of Human Decisions* (Macmillan, New York).

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

Clifford WK (2010) *The Ethics of Belief and Other Essays* (Prometheus Books, Amherst, NY).

Clough DR, Wu A (2022) Artificial intelligence, data-driven learning, and the decentralized structure of platform ecosystems. *Acad. Management Rev.* 47(1):184–189.

Coady CAJ (1992) *Testimony: A Philosophical Study* (Oxford University Press, Oxford, UK).

Constant A, Friston KJ, Clark A (2024) Cultivating creativity: Predictive brains and the enlightened room problem. *Philos. Trans. Roy. Soc. London B Biol. Sci.* 379(1895):20220415.

Cosmides L, Tooby J (2013) Evolutionary psychology: New perspectives on cognition and motivation. *Annual Rev. Psych.* 64:201–229.

Crouch TD (2002) *A Dream of Wings: Americans and the Airplane, 1875–1905* (WW Norton & Company, New York).

Csaszar FA, Levinthal DA (2016) Mental representation and the discovery of new strategies. *Strategic Management J.* 37(10): 2031–2049.

Csaszar FA, Steinberger T (2022) Organizations as artificial intelligences: The use of artificial intelligence analogies in organization theory. *Acad. Management Ann.* 16(1):1–37.

Csaszar FA, Ketkar HJ, Kim H (2024) AI and strategic decision making. Working paper, Bocconi University, Milan.

Dasgupta I, Schulz E, Tenenbaum JB, Gershman SJ (2020) A theory of learning to infer. *Psych. Rev.* 127(3):412–441.

Davenport TH, Kirby J (2016) *Only Humans Need Apply: Winners and Losers in the Age of Smart Machines* (Harper Business, New York).

Dewey J (1916) *Essays in Experimental Logic* (University of Chicago Press, Chicago).

Du N, Huang Y, Dai AM, Tong S, Lepikhin D, Xu Y, Cui C (2022) Glam: Efficient scaling of language models with mixture-of-experts. *Internat. Conf. Machine Learn.* (PMLR, New York), 5547–5569.

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 J.* 43(7):1287–1318.

Ehrig T, Felin T, Zenger T (2024) Causal reasoning and the scientific entrepreneur: Beyond Bayes. Agrawal A, Camuffo A, Gambardella A, Gans J, Scott E, Stern S, eds. *Bayesian Entrepreneurship* (MIT Press, Cambridge, MA).

Eisfeldt AL, Schubert G (2024) AI and finance. Preprint, submitted October 15, https://dx.doi.org/10.2139/ssrn.4988553.

Feigenbaum EA (1963) Artificial intelligence research. *IEEE Trans. Inform. Theory* 9(4):248–253.

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

Felin T, Koenderink J (2022) A generative view of rationality and growing awareness. *Frontiers Psych.* 13:807261.

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

Felin T, Gambardella A, Zenger T (2021a) *Value Lab: A Tool for Entrepreneurial Strategy* (Management & Business Review, Columbia, ML).

Felin T, Kauffman S, Zenger T (2023) Resource origins and search. *Strategic Management J.* 44(6):1514–1533.

Felin T, Koenderink J, Krueger JI (2017) Rationality, perception, and the all-seeing eye. *Psychonomic Bull. Rev.* 24:1040–1059.

Felin T, Gambardella A, Novelli E, Zenger T (2024) A scientific method for startups. *J. Management* 50(8):3080–3104.

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

Felin T, Koenderink J, Krueger JI, Noble D, Ellis GF (2021b) The data-hypothesis relationship. *Genome Biol.* 22(1):1–6.

Festinger L, Riecken HW, Schachter S (1956) *When Prophecy Fails* (University of Minnesota Press, Minneapolis).

Franceschelli G, Musolesi M (2023) On the creativity of large language models. Preprint, submitted March 27, https://arxiv.org/abs/2304.00008.

Friedman JH, Popescu BE (2008) Predictive learning via rule ensembles. *Ann. Appl. Statist.* 2(3):916–954.

Friston K, Kiebel S (2009) Predictive coding under the free-energy principle. *Philos. Trans. Roy. Soc. London B Biol. Sci.* 364(1521): 1211–1221.

Friston K, Da Costa L, Tschantz A, Kiefer A, Salvatori T, Neacsu V, Buckley CL (2023) Supervised structure learning. Preprint, submitted November 17, https://arxiv.org/abs/2311.10300.

Gagliardi L (2024) The role of cognitive biases in conspiracy beliefs: A literature review. *J. Econom. Surveys.* Forthcoming.

Gans JS, Stern S, Wu J (2019) Foundations of entrepreneurial strategy. *Strategic Management J.* 40(5):736–756.

Gavetti G, Levinthal D (2000) Looking forward and looking backward: Cognitive and experiential search. *Admin. Sci. Quart.* 45(1):113–137.

Gennaioli N, Shleifer A (2018) *A Crisis of Beliefs: Investor Psychology and Financial Fragility* (Princeton University Press, Princeton, NJ).

Gershman SJ, Horvitz EJ, Tenenbaum JB (2015) Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* 349(6245):273–278.

Gigerenzer G, Goldstein DG (2024) Herbert Simon on the birth of the mind-computer metaphor. Augier M, March JG, eds. *Elgar Companion to Herbert Simon* (Edward Elgar Publishing, Cheltenham, UK).

Gilkerson J, Richards JA, Warren SF, Montgomery JK, Greenwood CR, Kimbrough Oller D, Paul TD (2017) Mapping the early language environment using all-day recordings and automated analysis. *Amer. J. Speech Language Pathology* 26(2):248–265.

Gingerich O, MacLachlan J (2005) *Nicolaus Copernicus: Making the Earth a Planet* (Oxford University Press, Oxford, UK).

Girotra K, Meincke L, Terwiesch C, Ulrich KT (2023) Ideas are dimes a dozen: Large language models for idea generation in innovation. Preprint, submitted August 2, https://dx.doi.org/10.2139/ssrn.4526071.

Goldman AI (1999) *Knowledge in a Social World* (Oxford University Press, New York).

Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning* (MIT Press, Cambridge, MA).

Goyal A, Bengio Y (2022) Inductive biases for deep learning of higher-level cognition. *Proc. Roy. Soc. London A* 478(2266):20210068.

Grace K, Stewart H, Sandkühler JF, Thomas S, Weinstein-Raun B, Brauner J (2024) Thousands of AI authors on the future of AI. Preprint, submitted January 5, https://arxiv.org/abs/2401.02843.

Gregory RW, Henfridsson O, Kaganer E, Kyriakou H (2021) The role of artificial intelligence and data network effects for creating user value. *Acad. Management Rev.* 46(3):534–551.

Griffin D, Tversky A (1992) The weighing of evidence and the determinants of confidence. *Cognitive Psych.* 24(3):411–435.

Griffiths TL, Chater N, Kemp C, Perfors A, Tenenbaum JB (2010) Probabilistic models of cognition: Exploring representations and inductive biases. *Trends Cognitive Sci.* 14(8):357–364.

Hahn U, Harris AJ (2014) What does it mean to be biased: Motivated reasoning and rationality. Ross BH, ed. *Psychology of Learning and Motivation*, vol. 61 (Academic Press, San Diego), 41–102.

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

Halliday MAK (1989) *Spoken and Written Language* (Oxford University Press, Oxford, UK).

Hart B, Risley TR (2003) The early catastrophe: The 30 million word gap by age 3. *Amer. Ed.* 27(1):4–9.

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

Hebb DO (1949) *The Organization of Behavior: A Neuropsychological Theory* (Psychology Press, New York).

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

Hinton GE (1992) How neural networks learn from experience. *Sci. Amer.* 267(3):144–151.

Hinton GE (2023) Will digital intelligence replace biological intelligence? *University of Toronto Lecture* (October 27), https://youtu.be/iHCeAotHZa4.

Hohwy J (2013) *The Predictive Mind* (OUP Oxford, Oxford, UK).

Hohwy J (2020) New directions in predictive processing. *Mind Language* 35(2):209–223.

Holmström J, Holweg M, Lawson B, Pil FK, Wagner SM (2019) The digitalization of operations and supply chain management: Theoretical and methodological implications. *J. Oper. Management* 65(8):728–734.

Holtzman A, Buys J, Du L, Forbes M, Choi Y (2019) The curious case of neural text degeneration. Preprint, submitted April 22, https://arxiv.org/abs/1904.09751.

Hong P, Ghosal D, Majumder N, Aditya S, Mihalcea R, Poria S (2024) Evaluating LLMs' mathematical competency through ontology-guided perturbations. Preprint, submitted January 17, https://arxiv.org/abs/2401.09395.

Jaeger J, Riedl A, Djedovic A, Vervaeke J, Walsh D (2024) Naturalizing relevance realization: Why agency and cognition are fundamentally not computational. Working paper, University of Vienna, Wien, Austria.

James W (1967) *The Writings of William James: A Comprehensive Edition* (University of Chicago Press, Chicago).

Jia N, Luo X, Fang Z, Liao C (2024) When and how artificial intelligence augments employee creativity. *Acad. Management J.* 67(1):5–32.

Johnson-Laird PN (1983) *Mental Models: Toward a Cognitive Science of Language, Inference, and Consciousness* (Harvard University Press, Cambridge, MA).

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

Kahneman D (2003) Maps of bounded rationality: Psychology for behavioral economics. *Amer. Econom. Rev.* 93(5):1449–1475.

Kahneman D (2011) *Thinking Fast and Slow* (Farrar, Straus & Giroux, New York).

Kahneman D (2018) A comment on artificial intelligence and behavioral economics. Agrawal A, Gans J, Goldfarb A, eds. *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press, Chicago), 608–610.

Kahneman D, Sibony O, Sunstein CR (2021) *Noise: A Flaw in Human Judgment* (Hachette Publishing, New York).

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

Kemp A (2023) Competitive advantages through artificial intelligence: Toward a theory of situated AI. *Acad. Management Rev.* 49(3):618–635.

Keynes JM (1921) *A Treatise on Probability* (Macmillan, London).

Kıcıman E, Ness R, Sharma A, Tan C (2023) Causal reasoning and large language models: Opening a new frontier for causality. Preprint, submitted April 28, https://arxiv.org/abs/2305.00050.

Kim H, Glaeser EL, Hillis A, Kominers SD, Luca M (2024) Decision authority and the returns to algorithms. *Strategic Management J.* 45(4):619–648.

Knill DC, Pouget A (2004) The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends Neuroscience* 27(12):712–719.

Kotseruba I, Tsotsos JK (2020) 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artificial Intelligence Rev.* 53(1):17–94.

Kralik JD, Lee JH, Rosenbloom PS, Jackson PC Jr, Epstein SL, Romero OJ, McGreggor K (2018) Metacognition for a common model of cognition. *Procedia Comput. Sci.* 145:730–739.

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Pereira F, Burges CJ, Bottou L, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 25 (Curran Associates Inc., Red Hook, NY).

Kruglanski AW, Jasko K, Friston K (2020) All thinking is "wishful" thinking. *Trends Cognitive Sci.* 24(6):413–424.

Kunda Z (1990) The case for motivated reasoning. *Psych. Bull.* 108(3):480–498.

Kvam PD, Pleskac TJ (2016) Strength and weight: The determinants of choice and confidence. *Cognition* 152:170–180.

Laird JE, Lebiere C, Rosenbloom PS (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.

Laird JE, Newell A, Rosenbloom PS (1987) SOAR: An architecture for general intelligence. *Artificial Intelligence* 33(1):1–64.

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

Lakhotia K, Kharitonov E, Hsu WN, Adi Y, Polyak A, Bolte B, Dupoux E (2021) On generative spoken language modeling from raw audio. *Trans. Assoc. Comput. Linguistics* 9:1336–1354.

Lansdell BJ, Kording KP (2019) Toward learning-to-learn. *Current Opinion Behav. Sci.* 29:45–50.

LeConte J (1888) The problem of a flying machine. *Sci. Monthly* 34:69–77.

LeCun Y (2017) A path to AI. *Future of Life Institute Lecture* (January), https://www.youtube.com/watch?v=bub58oYJTm0.

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 Machines* 17:391–444.

Levin M (2024) AI: A bridge toward diverse intelligence and humanity's future. Working paper, Tufts University, Medford, MA.

Lewin K (1943) Psychology and the process of group living. *J. Soc. Psych.* 17(1):113–131.

Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Kiela D (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, vol. 33 (Curran Associates Inc., Red Hook, NY), 9459–9474.

Lu C, Lange RT, Foerster J, Clune J, Ha D (2024) The AI scientist: Toward fully automated open-ended scientific discovery. Preprint, submitted August 12, https://arxiv.org/abs/2408.06292.

Mahowald K, Ivanova AA, Blank IA, Kanwisher N, Tenenbaum JB, Fedorenko E (2024) Dissociating language and thought in large language models. *Trends Cognitive Sci.* 28(6):517–540.

Manning BS, Zhu K, Horton JJ (2024) Automated social science: Language models as scientist and subjects. NBER Working Paper No. 32381, National Bureau of Economic Research, Cambridge, MA.

Marr D (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (MIT Press, Cambridge, MA).

Mastrogiorgio A, Felin T, Kauffman S, Mastrogiorgio M (2022) More thumbs than rules: Is rationality an exaptation? *Frontiers Psych.* 13:805743.

McCarthy J, Hayes PJ (1969) Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence* 4:463–502.

McCarthy J, Minsky ML, Rochester N, Shannon CE (2007) A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine* 27(4):12.

McClelland JL, Rumelhart DE (1981) An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psych. Rev.* 88(5):375–407.

McCorduck P (2004) *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence* (CRC Press, Boca Raton, FL).

McCoy RT, 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. *Trans. Assoc. Comput. Linguistics* 11:652–670.

McCoy RT, Yao S, Friedman D, Hardy MD, Griffiths TL (2024) Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proc. Natl. Acad. Sci.* 121(41):e2322420121.

McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophysics* 5:115–133.

McCullough D (2015) *The Wright Brothers* (Simon and Schuster, New York).

Melville GW (1901) The engineer and the problem of aerial navigation. *North Amer. Rev.* 173(541):820–831.

Merdes C, Von Sydow M, Hahn U (2021) Formal models of source reliability. *Synthese* 198:5773–5801.

Minaee S, Mikolov T, Nikzad N, Chenaghlu M, Socher R, Amatriain X, Gao J (2024) Large language models: A survey. Preprint, submitted February 9, https://arxiv.org/abs/2402.06196.

Morris MR, Sohldickstein J, Fiedel N, Warkentin T, Dafoe A, Faust A, Legg S (2023) Levels of AGI: Operationalizing progress on the path to AGI. Preprint, submitted November 4, https://arxiv.org/abs/2311.02462.

Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, Mian A (2023) A comprehensive overview of large language models. Preprint, submitted July 12, https://arxiv.org/abs/2307.06435.

Newcomb S (1901) Is the airship coming? *McClure's Magazine* 17:432–435.

Newell A (1990) *Unified Theories of Cognition* (Harvard University Press, Cambridge, MA).

Newell A, Simon HA (1963) GPS, a program that simulates human thought. Feigenbaum EA, Feldman J, eds. *Comput. Thought* (McGraw-Hill, New York), 279–293.

Novelli E, Spina C (2022) When do entrepreneurs benefit from acting like scientists? A field experiment in the UK. Preprint, submitted July 28, 2021, https://dx.doi.org/10.2139/ssrn.3894831.

Parr T, Friston KJ (2017) Working memory, attention, and salience in active inference. *Sci. Rep.* 7(1):14678.

Pearl J, Mackenzie D (2018) *The Book of Why: The New Science of Cause and Effect* (Basic Books, New York).

Peirce CS (1957) The logic of abduction. Thomas V, ed. *Peirce's Essays in the Philosophy of Science* (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. *Biol. Psych.* 186:108741.

Pilgrim C, Sanborn A, Malthouse E, Hills TT (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., Cambridge, MA).

Pinker S (2021) *Rationality: What It Is, Why It Seems Scarce, Why It Matters* (Penguin, New York).

Polanyi M (1958) *Personal Knowledge* (The University of Chicago Press, Chicago).

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, Netherlands), 57–71.

Poldrack RA (2021) The physics of representation. *Synthese* 199(1–2): 1307–1325.

Popper K (1991) *All Life Is Problem Solving* (Routledge, London).

Puranam P, Stieglitz N, Osman M, Pillutla MM (2015) Modelling bounded rationality in organizations: Progress and prospects. *Acad. Management Ann.* 9(1):337–392.

Raisch S, Fomina K (2024) Combining human and artificial intelligence: Hybrid problem-solving in organizations. *Acad. Management Rev.* Forthcoming.

Ramsey FP (1931) *The Foundations of Mathematics and Other Logical Essays* (Cambridge University Press, Cambridge, UK).

Rao H, Greve HR (2024) The plot thickens: A sociology of conspiracy theories. *Annual Rev. Sociol.* 50:191–207.

Rescorla M (2015) The computational theory of mind. Zalta EN, ed. *The Stanford Encyclopedia of Philosophy*, Winter 2015 ed. (Metaphysics Research Lab, Stanford University, Stanford, CA).

Resnik P (2024) Large language models are biased because they are large language models. Preprint, submitted June 19, https://arxiv.org/abs/2406.13138.

Riedl R (1984) *Biology of Knowledge: The Evolutionary Basis of Reason* (Wiley, New York).

Roli A, Jaeger J, Kauffman SA (2022) How organisms come to know the world: Fundamental limits on artificial general intelligence. *Frontiers Ecology Evolution* 9:1035–1050.

Rosenblatt F (1958) The perceptron: A probabilistic model for information storage and organization in the brain. *Psych. Rev.* 65(6):386–408.

Rosenblatt F (1962) *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms* (Spartan Books, Washington, DC).

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 DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. *Nature* 323(6088):533–536.

Russell SJ, Norvig P (2022) *Artificial Intelligence: A Modern Approach* (Pearson, London).

Scheffer M, Borsboom D, Nieuwenhuis S, Westley F (2022) Belief traps: Tackling the inertia of harmful beliefs. *Proc. Natl. Acad. Sci. USA* 119(32):e2203149119.

Schwöbel S, Kiebel S, Marković D (2018) Active inference, belief propagation, and the Bethe approximation. *Neural Comput.* 30(9):2530–2567.

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. Preprint, submitted January 23, https://arxiv.org/abs/1701.06538.

Shumailov I, Shumaylov Z, Zhao Y, Papernot N, Anderson R, Gal Y (2024) AI models collapse when trained on recursively generated data. *Nature* 631(8022):755–759.

Simon HA (1955) A behavioral model of rational choice. *Quart. J. Econom.* 69(1):99–118.

Simon HA (1980) Cognitive science: The newest science of the artificial. *Cognitive Sci.* 4(1):33–46.

Simon HA (1985a) Artificial intelligence: Current status and future potential. National Research Council Report, Office of Naval Research, Arlington, VA.

Simon HA (1985b) Human nature in politics: The dialogue of psychology with political science. *Amer. Political Sci. Rev.* 79(2):293–304.

Simon HA (1990) Invariants of human behavior. *Annual Rev. Psych.* 41(1):1–20.

Simon HA, Newell A (1958) Heuristic problem solving: The next advance in operations research. *Oper. Res.* 6(1):1–10.

Spelke ES, Breinlinger K, Macomber J, Jacobson K (1992) Origins of knowledge. *Psych. Rev.* 99(4):605–632.

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

Tannen D (2007) *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse* (Cambridge University Press, Cambridge, MA).

Tervo DGR, Tenenbaum JB, Gershman SJ (2016) Toward the neural implementation of structure learning. *Current Opinion Neurobiology* 37:99–105.

Tranchero M, Brenninkmeijer CF, Murugan A, Nagaraj A (2024) Theorizing with large language models. Preprint, submitted October 8, https://dx.doi.org/10.2139/ssrn.4978831.

Turing AM (1948/1992) Intelligent machinery. Ince DC, ed. *Mechanical Intelligence, Collected Works of A. M. Turing* (North Holland, Amsterdam), 107–127.

Turing AM (1950) Computing machinery and intelligence. *Mind* 59(236):433–460.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Polosukhin I (2017) Attention is all you need. *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates Inc., Red Hook, NY).

Wright O, Wright W (1881–1940) *Wilbur and Orville Wright papers*. Library of Congress. Accessed October 2023, 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 J.* 44(12):2922–2949.

Yanai I, Lercher M (2020) A hypothesis is a liability. *Genome Biol.* 21(1):1–5.

Yin H (2020) The crisis in neuroscience. Lowe R, Powers WT, Marken CS, eds. *The Interdisciplinary Handbook of Perceptual Control Theory* (Academic Press, Cambridge, MA), 23–48.

Yiu E, Kosoy E, Gopnik A (2023) Transmission vs. truth, imitation vs. innovation: What children can do that large language and language-and-vision models cannot (yet). *Perspect. Psych. Sci.* 19(5):874–883.

Zellweger T, Zenger T (2023) Entrepreneurs as scientists: A pragmatist alternative to the creation-discovery debate. *Acad. Management Rev.* 47(4):696–699.

Zhou HY, 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 Engrg.* 7:743–755.

Teppo Felin is the Douglas D. Anderson endowed professor at the Huntsman School of Business, Utah State University, as well as associate scholar at Saïd Business School, University of Oxford. His research is broadly focused on cognition, decision making, organizational economics, strategy, entrepreneurship, and innovation.

Matthias Holweg is the American Standard Companies professor of operations management at the Saïd Business School at the University of Oxford. His research explores how organizations can leverage digital technologies to create and capture value with particular focus on enhancing operational excellence through data-driven process improvements.