Perspective

Artificial intelligence and illusions of understanding in scientific research

https://doi.org/10.1038/s41586-024-07146-0

Received: 31 July 2023

Accepted: 31 January 2024

Published online: 6 March 2024

Check for updates

Lisa Messeri[1,4]✉ & M. J. Crockett[2,3,4]✉

Scientists are enthusiastically imagining ways in which artificial intelligence (AI) tools might improve research. Why are AI tools so attractive and what are the risks of implementing them across the research pipeline? Here we develop a taxonomy of scientists' visions for AI, observing that their appeal comes from promises to improve productivity and objectivity by overcoming human shortcomings. But proposed AI solutions can also exploit our cognitive limitations, making us vulnerable to illusions of understanding in which we believe we understand more about the world than we actually do. Such illusions obscure the scientific community's ability to see the formation of scientific monocultures, in which some types of methods, questions and viewpoints come to dominate alternative approaches, making science less innovative and more vulnerable to errors. The proliferation of AI tools in science risks introducing a phase of scientific enquiry in which we produce more but understand less. By analysing the appeal of these tools, we provide a framework for advancing discussions of responsible knowledge production in the age of AI.

Scientific futures that incorporate AI are being proposed in which 'self-driving' laboratories abound[1,2], human participants can be replaced by generative AI[3,4] and 'AI scientists' write research papers[5] and produce Nobel-prizewinning discoveries[6]. These visions go beyond positioning AI as a mere tool, instead positing autonomous collaborators that can overcome the limits that human capabilities currently impose on advancing science. Although such proposals might sound like science fiction, they are being published in prestigious scientific journals and actively pursued with the support of powerful institutions. Indeed, the National Academies of Sciences, Engineering, and Medicine recently held a workshop in which academic researchers, industry spokespeople and funders evaluated the potential of AI to become "an autonomous researcher performing scientific discovery" (www.nationalacademies.org/event/40455_10-2023_ai-for-scientific-discovery-a-workshop). What should we make of the enthusiasm for visions of science that cede ever greater autonomy to AI? We need to consider how the widespread adoption of AI tools might impact scientific knowledge production and understanding.

Researchers evaluating the risks of AI in science and society have recognized a variety of ethical concerns, including algorithmic bias[7–9], environmental costs[10,11], public misunderstanding of the capabilities of AI[12] and exploitative labour practices[11,13]. Attention has also been paid to AI's epistemic risks (risks of being wrong; Box 1) arising from errors and 'hallucinations'[10,14–16], failures of reproducibility[17,18] and lack of interpretability[16,19,20]. Scholars recognize that technical approaches alone are inadequate for addressing the ethical concerns of AI[21]. Yet there remains an optimism that the aforementioned epistemic risks will yield to exclusively technical solutions.

Here we focus on a set of under-discussed epistemic risks of AI in science that are likely to evade purely technical solutions (Box 2). Scientists interested in using AI in their research and researchers who study AI must evaluate these risks now, while AI applications are still nascent, because they will be much more difficult to address if AI tools become deeply embedded in the research pipeline[22]. To be clear, we do not take the position that AI should never be used in scientific research. Rather, we seek to identify risks that some, but not necessarily all, AI approaches might create in our pursuit of scientific understanding.

Our expertise in cognitive science, epistemology, anthropology and science and technology studies provides insights into how adopting AI in scientific research can impede scientific understanding despite promising to improve it. By considering why scientists are motivated to use AI tools, we develop a taxonomy of scientists' proposed visions for AI across the research pipeline and identify features of AI tools that make them attractive as knowledge-production partners. However, individuals who trust AI tools to overcome their own cognitive limitations become susceptible to illusions of understanding (Fig. 1 and Box 1) in which they believe that they understand more about the world than they actually do. Such illusions obscure the scientific community's ability to see the formation of scientific monocultures (Box 1) in which certain methods, questions and viewpoints come to dominate alternative approaches, making science less innovative and more vulnerable to errors. The proliferation of AI tools in science risks introducing a phase of scientific enquiry in which we produce more but understand less.

Visions of AI for scientific research

Scientists must contend with various research constraints, including finite time, fixed budgets and limited cognitive capabilities. AI tools are viewed as solutions to these barriers, enabling scientists to be more productive (in terms of generating more scientific work) and more objective (free from bias and subjectivity). We identify four distinct visions of AI: Oracle, Surrogate, Quant and Arbiter (capitalized

[1]Department of Anthropology, Yale University, New Haven, CT, USA. [2]Department of Psychology, Princeton University, Princeton, NJ, USA. [3]University Center for Human Values, Princeton University, Princeton, NJ, USA. [4]These authors contributed equally: Lisa Messeri, M. J. Crockett. ✉e-mail: lisa.messeri@yale.edu; mj.crockett@princeton.edu

Box 1

Glossary

Cognitive diversity means differences in how people think and approach a problem. This can come from disciplinary training, skills, career stage or neurodivergence.

Communities of knowledge are groups of individuals with distributed knowledge and understanding that allow individual community members to benefit from expertise held by others.

Demographic diversity means differences in race, class, gender, sexuality, culture, religion, ethnicity, ability, age and so on that often map on to different life experiences.

Epistemic risks refer to a broad class of risks arising from holding incorrect beliefs.

Epistemic trust refers to trusting agents in their capacity to provide accurate, reliable information.

The **illusion of explanatory depth** is an illusion of understanding in which someone incorrectly believes they have a deeper or more comprehensive level of understanding then they actually do (Fig. 1a).

The **illusion of exploratory breadth** is an illusion of understanding that accompanies monocultures of knowing in which scientists falsely believe they are exploring the full space of testable hypotheses, whereas they are actually exploring a narrower space of hypotheses testable using AI tools (Fig. 1b).

The **illusion of objectivity** is an illusion of understanding accompanying a monoculture of knowers in which scientists falsely believe that AI tools do not have a standpoint or are able to represent all possible standpoints, whereas AI tools actually embed the standpoints of their training data and developers (Fig. 1c).

Illusions of understanding are a class of metacognitive errors that arise from holding incorrect beliefs about the nature of one's understanding.

A **monoculture of knowers** emerges from prioritizing a particular standpoint and all of the values and assumptions it contains, and this influences the research questions that are asked and the way evidence is interpreted. Visions of AI for science reinforce the idea of a singular, authoritative, objective knower with an unbiased 'view from nowhere'.

Monocultures of knowing emerge from prioritizing one approach to asking research questions and determining when satisfactory understanding has been achieved, marginalizing alternative approaches. In the context of AI-driven research, a monoculture of knowing could arise from prioritizing the quantitative, reductive and predictive approaches that AI tools are designed to optimize.

Scientific monocultures arise when one approach to knowledge production becomes widely adopted at the expense of alternative approaches in a research domain, eliminating multiple forms of diversity from the process of knowledge production.

A **standpoint** refers to a socially situated perspective that provides a potential epistemic advantage, in the sense that individuals who occupy a particular standpoint may be in a better position to know and understand certain concepts than others who do not occupy that standpoint. In the context of scientific knowledge production, a standpoint can refer to perspectives arising from one's social position on the basis of factors including (but not limited to) race, ethnicity, gender, class, disability and/or relevant experiences such as background, training and career stage.

to signal that these visions are not inevitable, but emerge from collective human endeavours). Each vision promises in distinct ways to enhance the capabilities of human scientists across the research pipeline (Table 1). Although these interventions are often framed as preliminary, they reveal how scientists are imagining the long-term potential of AI-driven science.

Scientists who describe advances in AI methods over the past decade have tended to classify them into two broad categories: predictive AI tools, which analyse patterns in their training datasets to make predictions about new data; and generative AI tools, which generate new data and text on the basis of predictions from patterns observed in their training datasets (Box 2). The four visions we describe below apply (and intermix) predictive and generative AI towards different goals in scientific research. This is not an exhaustive list of AI visions in the scientific literature. We focus here only on the visions of AI that are most relevant to the epistemic risks that we address in more detail below.

AI as Oracle

At the start of the research pipeline, the vision of AI as an Oracle that can digest and communicate scientific knowledge promises to solve an important problem: the deluge of published material "threatening to exceed the cognitive limits of human processing capacities"[23,24]. The exponentially growing literature is of uneven quality[25], and this has been attributed to institutional pressures to "publish or perish"[26] and the rise of predatory publishers. Multiple AI tools that can query, digest and summarize the published record are being developed and evaluated[23,27,28]. Oracles are also positioned as solutions to the problem of generating new hypotheses from vast amounts of literature; for example, in applications of predictive and generative AI to protein folding[29–31] and materials science[32]. Although the enthusiasm for Oracles is largely shared by the authors of these publications, some researchers have observed how Oracles can detract from scientific understanding[33], a point we discuss below.

Oracles are thought to offer several "advantages over humans", including "increased precision and exhaustiveness" and reduced "researcher bias"[26]. AI is positioned to exceed human capabilities "by revealing hidden connections between findings"[34], even if such feats are not yet realized[35]. One computer scientist working at an AI research laboratory predicts a "tireless" Oracle in which "scientific research is mostly read by machines"[36].

AI as Surrogate

Data collection is time consuming and expensive. The vision of AI as Surrogate is one that can enhance a laboratory's measurement capabilities; for example, for projects where "traditional data collection is impractical"[4]. By leveraging the potential of generative AI as an "agent of replacement"[37], proposals for "silicon subjects"[38] and "synthetic data points"[2] span scientific disciplines. In the social sciences, some researchers are excited about the potential for generative AI to simulate human participants[3,4,38–41]. In the physical and biological sciences, proposed applications for generative AI include augmenting small datasets or simulating new data to study phenomena such as cosmological structures[42], medical images[43,44] and nucleic-acid sequences[45].

Surrogates are imagined to not only provide data that are difficult or expensive to obtain but also, in some cases, to exceed the quality of data collected without AI assistance. In the social sciences, Surrogates are positioned as ideal research participants that "can rapidly answer hundreds of questions without fatigue" and "need fewer incentives than humans to give reliable responses"[3]. If trained properly, generative AI tools are envisioned as representing "a vast array of human

Box 2

AI visions as sociotechnical visions

The visions of AI Oracles, Surrogates, Quants and Arbiters appear in papers authored both by scientists without AI expertise and by researchers who study AI, either through interdisciplinary collaborations (and industry partnerships) or from a single discipline. So far, these visions have drawn primarily from two broad classes of AI application. Predictive AI refers to AI applications that use machine learning methods to make classifications and/or predictions on the basis of patterns learned in a training dataset. Examples include deep neural network models for image classification and ensemble models for forecasting political events. Generative AI, which has been developed more recently, refers to AI applications that learn the distribution of data in a training dataset and can generate new samples from that distribution. Examples include generative diffusion models for image generation and large language models, such as OpenAI's GPT-4, Google's Gemini or Meta's LLaMa, that can generate text that mimics human language.

Given the pace of AI research, the techniques of tomorrow might look very different from those around today. For instance, researchers are actively working towards developing AI models that have human-like causal learning and reasoning capabilities[183–185] (although many hurdles remain before this is achieved[186]). Nevertheless, the epistemic risks we identify here, although motivated by AI visions based primarily on predictive and generative AI, are likely to persist even when AI techniques become more sophisticated and scientists incorporate them into ever-more-capable Oracles, Surrogates, Quants and Arbiters. Indeed, as AI tools become more convincingly human-like, the epistemic risks that arise from uncritically trusting such tools as scientific collaborators may even be amplified.

Understanding the inherently social nature of these epistemic risks underscores the inadequacy of purely technical solutions for addressing them. Illusions of understanding that arise from an overreliance on AI in science (Fig. 1) cannot be overcome by using more sophisticated AI models or by preventing errors such as hallucinations. Rather, they require sociotechnical approaches that account for the inseparability of social and technical dynamics[21,174,187]. In other words, because scientific research is a fundamentally social process[147,148], evaluating the epistemic risks of AI for science requires not only technical assessments, but also an understanding of the social and cognitive processes through which scientists extend epistemic trust, decide what research questions to pursue and interpret the results of experiments.

experiences and perspectives … offering a more accurate portrayal of human behaviour and social dynamics than those from conventional methods"[4]. In the physical and biological sciences, data augmentation is reported to "enhance model robustness and generalizability" and the "diversity of the data"[2].

AI as Quant

As many researchers have observed, with both celebration[46] and concern[47], the paradigm of 'big data' has influenced disciplines across the sciences, emphasizing computational approaches. The vast datasets yielded by big data approaches pose challenges to human capabilities that are not easily addressed by established statistical approaches, either in terms of the volume of data that need to be curated or the potential complexity of the models needed to explain and predict the data[48,49]. Quants offer solutions for both data preparation and analysis. In biology, predictive AI tools are already being used for the automated annotation of protein function[50,51] and cell types[52]. In the social sciences, generative AI tools are being discussed as solutions to annotating and even ascribing meaning to text, images and qualitative data[53,54], tasks that previously required substantial human labour. For analysing large and complex datasets, Quants are imagined that "can extract meaningful representations of scientific data", retaining "as much information about the data as possible while remaining simple and accessible"[2].

There is excitement about the potential of AI to open up new knowledge "frontiers"[35]; for example, in mathematics[55] and in "uncovering new cognitive and behavioural phenomena"[56]. Quants might discover features in natural data that are predictively valid but cognitively inaccessible to human minds[48,57,58]. Recognizing that many AI models are too complex for human scientists to understand, there is also discussion of AI being used to produce simplified model explanations that are comprehensible to human minds but still retain fidelity to the original model[59].

AI as Arbiter

Whereas Oracles are proposed as solutions for information overload during hypothesis generation, Arbiters are seen as responses to the same problem at the end of the research pipeline. New tools are required to address "the constant growth of submission volume"[60] that editors and peer reviewers face. The growing labour demands for the peer review of grants and papers have thus spurred researchers to discuss "the potential for artificial intelligence to replace peer review"[61]. Arbiters have been offered to assist with the preliminary screening of submitted manuscripts[62,63], and generative AI is imagined as being able to write reviews[64]. Arbiters are also proposed to intervene in the replication crisis that has plagued fields in which (time- and cost-intensive) experimental replication has been difficult[65]. Predictive AI approaches that "require little subjective peer judgment and minimize costs" are proposed to provide "systematic, fast, and accurate"[66] predictions of the reproducibility of scientific findings[67] or even entire subfields[68].

As well as enhancing efficiency, AI is proposed to be an authoritative judge, removing human subjectivity and bias from contentious decision-making processes. Indeed, researchers have suggested that the process of developing AI techniques for peer review can "uncover biases in [human] decision-making"[60] that have historically favoured certain scholars and institutions. Although it is cautioned that at present AI should not be used to fully replace human experts[60], the vision of Arbiters suggests a future in which this caution will no longer be warranted.

Summary

Distinct AI interventions are being proposed that span the entire research pipeline. Each vision identifies different problems and solutions, but there are also interdependencies, such that as one vision gets adopted, the uptake of subsequent visions is required or enhanced. For example, the large datasets that Surrogates produce require Quants to analyse them. Furthermore, the efficiency promised by Surrogates and Quants will yield even more publications for Arbiters to adjudicate and for Oracles to digest and summarize. These visions also reinforce one another by converging on two broad goals: to enhance scientific productivity by overcoming scientists' limited time, attention and cognitive capacities; and to enhance scientific objectivity by overcoming

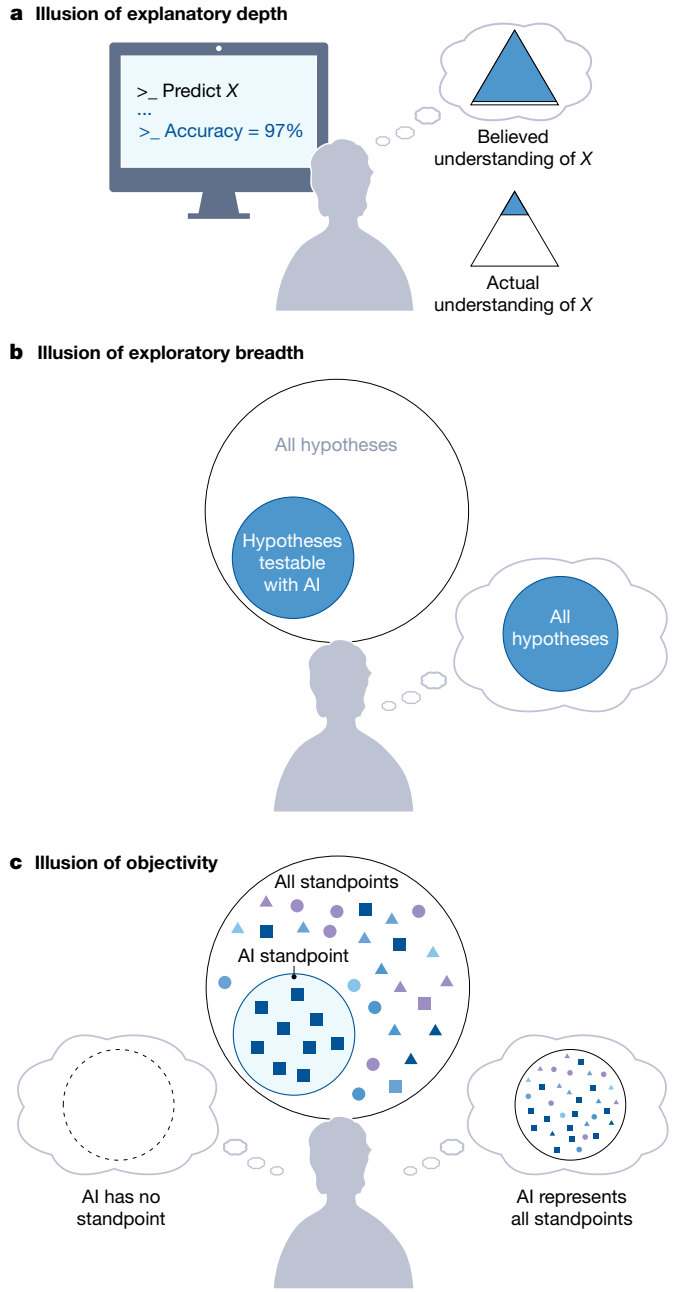

Fig. 1 | Illusions of understanding in AI-driven scientific research. a, Scientists using AI tools for their research may experience an illusion of explanatory depth. In this example, a scientist uses an AI Quant to model a phenomenon (*X*) and believes they understand *X* with more depth than they actually do. **b**, In a monoculture of knowing, scientists are vulnerable to an illusion of exploratory breadth, in which they falsely believe they are exploring a space of all testable hypotheses, whereas they are actually exploring a narrower space of hypotheses that are testable with AI tools. **c**, In a monoculture of knowers, scientists are vulnerable to an illusion of objectivity, in which they falsely believe that AI tools do not have a standpoint (as desired for Oracles and Arbiters) or are able to represent all possible standpoints (as desired for Surrogates in research using human participants), whereas AI tools actually embed the standpoints of their training data and their developers.

scientists' subjectivity and bias. Achieving these goals is expected to improve scientific understanding.

These potential benefits of AI are worth taking seriously, but it is critical that scientists and developers of AI tools also consider an alternative possibility: that under some conditions, AI tools may in fact limit, rather than enhance, scientific understanding. In other words, alongside their potential epistemic benefits, AI Oracles, Surrogates, Quants and Arbiters carry epistemic risks when scientists trust them as knowledge-production partners.

Epistemic trust in AI tools

Why do scientists trust AI tools? In this section, we explain how certain features of scientific work, the AI tools themselves and the solutions they offer make them likely to be treated not just as tools but perhaps even as fully fledged collaborators that are part of the research team. This creates epistemic risks beyond those that scientists contend with when using computational tools that are not treated as partners in knowledge production. Specifically, treating AI tools as scientific collaborators makes scientists vulnerable to illusions of understanding, a class of metacognitive errors that arise from holding incorrect beliefs about the nature of one's understanding[69,70] (Fig. 1).

Including AI in communities of knowledge

Individual human minds are limited in that people cannot process all the information available to them in practical or scientific deliberation. However, we can greatly expand the range of our understanding by dividing cognitive labour among the members of a group[71–74]. In communities of knowledge (Box 1), individuals benefit from the expertise held by other trusted individuals[69,75–78]. To benefit from communities of knowledge, people need to track how knowledge clusters in other minds[75,79] and determine who deserves epistemic trust[69,77,78,80] in order to consult and defer to them for advice[81].

Increasingly, people are granting epistemic trust not only to people, but also to digital tools, including AI[82–84]. Experimental studies show that people are more likely to trust AI for tasks that share key features with scientific research: tasks that can be described as objective (as opposed to subjective)[85,86], that require deliberation (as opposed to intuition)[87], where human biases are undesirable[88,89] and under conditions of high workloads[90] and increasing task difficulty[91]. This work suggests that scientific research—which demands objectivity, advanced analytic thinking and productivity—is a setting ripe for trust in AI. Indeed, the perceived cognitive and material limitations of scientists, and the hope, outlined in the visions above, that AI tools can transcend these limitations, makes trusting AI tools and welcoming them into our communities of knowledge deeply appealing.

AI Oracles, Surrogates, Quants and Arbiters are often described as not just mere tools, but as anthropomorphized "partners for scientists"[59] or "scientific assistant[s]"[35]. These AI visions are praised for overcoming human limitations and are thus more specifically anthropomorphized as 'superhuman' in ways that are likely to enhance epistemic trust. In particular, the visions reveal optimism that AI tools can conduct research more objectively than human scientists can, as seen in discussions of Oracles being less likely to cherry-pick the literature to support their desired hypotheses[26] and of Arbiters being less likely to show favouritism in scientific peer review[60]. Positioning AI tools in this way is likely to enhance epistemic trust because perceived objectivity is an important marker of credibility[92–94].

Furthermore, experiments have demonstrated that people extend epistemic trust to those perceived to have deeper understanding of the target phenomenon[81]. Although the question of what AI models can 'understand' remains controversial[14], the AI visions (alongside the language of 'deep learning', 'deep neural networks' and marketing hype[95]) portray AI as having a depth of understanding beyond what can be grasped by limited human minds. This portrayal could make AI tools seem more credible than human experts.

Finally, AI tools might seem especially trustworthy because they offer solutions that share features with satisfying explanations. Experiments show people prefer answers and explanations that are simple[96,97], broad[96,98,99], reductive[100–102] and quantitative[103,104]. These qualities also

Table 1 | Visions of AI across the research pipeline

Vision	Research stage	Limits to overcome	Vision
AI as Oracle	Study design	There is too much literature to digest; scientific publications vary in quality; readers are biased; too many research paths to choose from	Tools that objectively and efficiently search, evaluate and summarize scientific literature and generate new hypotheses
AI as Surrogate	Data collection	Data are too difficult, time consuming or expensive to obtain	Tools that accurately and tractably generate surrogate data points from natural complex systems, including human participants
AI as Quant	Data analysis	Data are too large or complex to curate and analyse	Tools that surpass the limits of human intellect in curating and analysing vast and complex datasets to produce new knowledge
AI as Arbiter	Peer review	There are too many papers and proposals to review; reviewers are biased	Tools that objectively and efficiently evaluate scientific merit and the replicability of findings

We derived this typology by analysing recent publications concerning the potential of AI to improve knowledge production across scientific disciplines. Included papers either used the general phrase 'artificial intelligence' or referred to specific methods classed under AI, most frequently machine learning, natural language processing and large language models. This table summarizes how the visions are responsive to different research stages, as well as the perceived limits to scientific capacity and efficiency.

feature prominently in scientists' visions of AI. Oracles are positioned as providing simplifying summaries of entire literatures; Quants are built to provide quantitative models of complex natural phenomena; Surrogates are viewed as representing the full breadth of humanity; and Arbiters are proposed to evaluate scientific work on reductive metrics such as 'quality' or 'replicability'. Although reductive and quantitative explanations tend to produce feelings of understanding, such feelings are not always correlated with actual understanding[100,105]. This can lead to illusions of understanding, which we consider next.

Illusions of understanding in communities of knowledge

Communities of knowledge offer clear epistemic benefits, enabling individuals to access much more knowledge than they could achieve alone. However, they also create epistemic risks. A well-established class of epistemic risk is illusions of understanding, in which individuals in communities of knowledge believe they understand more than they actually do. One way this manifests is when individuals mistake other community members' understanding for their own. For example, when online participants learned that a scientific phenomenon was well understood by scientists, they reported greater understanding of that phenomenon themselves, even when they had no basis for actual understanding[77]. Importantly, such illusory understanding is also observed when cognitive labour is offloaded to machines. Several studies have demonstrated that people overestimate what they know and understand when they search the internet for explanations, mistaking online access to information for their own personal knowledge[83,106,107]. Other work has shown that students overestimate their own knowledge when they answer test questions with AI assistance[108].

Illusions of understanding are particularly pronounced for explanatory knowledge relative to other kinds of knowledge, such as facts, procedures or narratives; this phenomenon is known as the illusion of explanatory depth[70] (Fig. 1a). Recent work suggests that this illusion inflates user understandings of AI systems[109]. Because explanatory knowledge is foundational to scientific understanding[110], scientists, like others in knowledge communities, are prone to the illusion of explanatory depth, overestimating the detail, coherence and depth of their scientific understanding[111]. One example is the 'prediction–explanation fallacy', in which scientists uncritically use prediction-optimized models for explanatory purposes, failing to appreciate that even the most accurate predictive model may bear little relation to the actual data-generating process[112–115]. Such risks can be compounded when scientists use AI tools outside their domain of expertise (a setting that, incidentally, has been demonstrated to increase trust in AI[116]), increasing the likelihood of errors because the users lack the expertise to know when the results are too good to be true[117]. There is evidence for considerable overoptimism in scientific claims that are based on machine learning model performance[17], and this probably arises from a poor understanding of the limits of machine prediction in fields beyond computer science.

The illusions of understanding we have reviewed so far are not unique to the AI context. Some relevant debates, such as the tension between prediction and explanation, long predate recent advances in AI[118]. In the next section, we will examine additional illusions of understanding that we anticipate emerging from a widespread proliferation of AI in science.

Epistemic risks of scientific monocultures

Because AI tools seem trustworthy and promise to enhance the quality and quantity of research, it is likely that research that relies on these tools will proliferate. Accordingly, several recent studies indicate increasing references to AI in publications and patents[119,120], and papers that use AI tools are cited more within and outside their disciplines[119]. In this section, we describe the epistemic risks that may arise if these trends continue and AI-assisted research comes to dominate the production of scientific knowledge.

To illustrate the nature of these risks, we draw on the metaphor of a monoculture. In agriculture, monoculture is the practice of growing only one crop species in a field at a time. This practice improves efficiency but makes the crop more vulnerable to pests and disease. We suggest that the efficiencies offered by AI tools can foster the growth of scientific monocultures, in which certain forms of knowledge production come to dominate all the rest. They can do so in two distinct but complementary ways: first, by prioritizing the kinds of questions and methods that are best suited for AI assistance (monocultures of knowing); and second, by prioritizing the types of standpoint that AI is able to express (monocultures of knowers). Just as plant monocultures are more vulnerable to pests and disease, scientific monocultures make our understanding of the world more vulnerable to error, bias and missed opportunities for innovation[121–124]. New tools and techniques are always prone to creating monocultures when scientists rush to exploit their benefits. However, the breadth of scientific applications predicted for AI tools[2,4] and their potential for inclusion in communities of knowledge as superhuman collaborators make the risks of AI-seeded monocultures particularly pernicious.

As well as threatening the robustness of science, monocultures of knowing and knowers create their own illusions of understanding (Fig. 1b,c). In these illusions, scientists incorrectly believe that AI tools advance scientific understanding, failing to appreciate that these tools instead narrow the scope of scientific knowledge production. Raising awareness of the epistemic risks of scientific monocultures, and their corresponding illusions of understanding, is a crucial step towards building systems of knowledge production that mitigate these risks.

Monocultures of knowing

Knowledge production, in terms of the questions asked and answers discovered, is strongly dependent on the methods and tools

available[47,125,126]. When there are incentives to use some methods over others, because they are cheaper, more efficient or more prestigious, those methods will proliferate. Consequently, the questions and answers afforded by those methods dominate knowledge production. For instance, when online studies became more popular in social science, self-report measures (which are easily deployed online) became more common than behavioural measures[127–129]. This is an example of a trend towards a monoculture of knowing, in which the 'crops' of research we produce are less diverse in terms of the questions they are able to ask and answer. Which ways of knowing are prioritized, and deprioritized, by an overreliance on AI tools in science?

One particularly enticing feature of AI tools is their promise to transform diverse, complex types of data, including human language, into the interoperable language of quantification. A salient example is the AI Surrogate that is proposed to simulate human research participants in silico[3,4,38]. Quantitative ways of knowing make information portable, aggregatable and comparable across contexts, which are essential activities when seeking to produce generalizable scientific knowledge[130,131]. However, quantitative ways of knowing strip out the contextual sensitivity and local details that are preserved by qualitative approaches. Indeed, eliminating these qualities lend quantitative approaches their powerful breadth but also make them "blunt instruments"[105]. Echoing concerns about the rise of big data as a research tool[47], research questions that require attention to subjectivity, subtlety and nuance, or otherwise cannot be quantified by a machine, are less likely to benefit from AI approaches and risk being marginalized. Such questions are especially common in the human sciences, in which visions of Surrogates proliferate, suggesting that these fields may be particularly vulnerable to distorted understandings yielded by an overreliance on quantitative ways of knowing.

To translate complex natural phenomena into quantitative models, researchers who use AI tools must often make simplifying approximations. This approach is common in computational social science, in which Quants are used to estimate proxies for concepts that are not directly observable[132,133], such as people's political ideology[134] or emotional expressions[135,136]. Tools that yield discoveries and high-impact publications are adopted by other teams that wish to study similar questions but lack the resources to build tools of their own, creating the potential for monocultures. Epistemic risks arise here from failing to appreciate the largely invisible "researcher degrees of freedom" that are embedded in the process of approximation[113]. Curating training sets and designing a training regime involves dozens or even hundreds of judgements that can imbue algorithms with the values of their creators[47,137]. These specific, often disciplinarily determined, decisions along the research pipeline can cause different researchers to arrive at different conclusions from the same starting data[115,138]. But once the algorithm is built, these choices disappear behind the objective sheen of a quantitative model[47,131,139], especially to those who did not build the model themselves. Failing to appreciate the decisions made early in tool design can lead to an overconfidence that established tools are the best or only way to model a particular phenomenon, foreclosing the discovery of alternative understandings of the approximated phenomenon that could expose errors or yield new discoveries[121].

Enthusiasm for AI tools is particularly concentrated around their predictive abilities, such as the envisioned ability of Oracles to predict complex biological structures[2] or of Arbiters to determine which studies are likely to be reproducible[66–68]. Some scientists have embraced the predictive prowess of machine learning approaches, arguing that we ought to prioritize predictive accuracy over developing accurate causal explanations[49]. This attitude is reflected in trends towards favouring prediction over explanation in science afforded by advances in AI[115,140]. However, others have highlighted epistemic risks inherent to predictive science[17,33,113,141,142]. Many of these risks are not unique to AI models, but as predictive AI methods become more complex, the associated risks become more challenging to articulate and address. Predictive AI models often lack interpretability[16,19,112], a property that is important for advancing scientific theories[143] and for adjusting behaviours when a model deems them insufficient[20,144]. And at the systemic level, aggregating knowledge across many models that prioritize prediction over explanation will make it much more difficult to identify the source of errors in those models, or even to recognize that there are errors at all[145]. This kind of systemic epistemic risk might be especially likely to accrue when new AI models are trained on the synthetic data outputs of previous models[146].

Given the distinct strengths and limitations that different ways of knowing bring to scientific research, knowledge production systems that lack diversity in ways of knowing will be more vulnerable to errors and missed opportunities. The epistemic trust placed in AI tools further compounds these risks because scientists might fail to recognize the elimination of diverse ways of knowing. The proliferation of AI tools makes scientists vulnerable to an illusion of exploratory breadth (Fig. 1b), whereby the subset of hypotheses that AI tools are good at evaluating are mistakenly viewed as being the entire set of testable hypotheses. Distracted by a deluge of new findings produced by AI tools, we may fail to appreciate that our search space has narrowed.

Monocultures of knowers

In their visions of AI, scientists anticipate a future in which biases of training sets and models have been overcome and AI tools are more reliable contributors to scientific projects than are human scientists. Oracles and Arbiters have eliminated subjective judgement from the tasks of evaluating the scientific record and assessing the merit of research. The synthetic data points of Surrogates make it possible to simulate a vast array of human experiences and perspectives. And the unbridled analytic capacity of Quants surpasses what any human knower alone could achieve. These AI tools are thus viewed as being objective and universal, replacing the work of diverse knowers and thus cultivating a monoculture of knowers. The epistemic risk of this monoculture becomes clear when considering what is lost when removing human diversity from scientific work, in the form of both demographic diversity (and different attendant life experiences) and cognitive diversity (arising from different disciplinary training, skills and problem-solving strategies) (Box 1).

Objectivity is widely held to be a core value of science. However, it is difficult to achieve in practice because scientific knowledge production is fundamentally a social endeavour[147,148]. Scientists have different standpoints (Box 1) that influence which questions they choose to pursue, how they ask these questions, what they take to be acceptable answers and how they frame the broader implications of those answers for future research[149–151]. Historically, the influence of standpoints on scientific knowledge production was invisible because scientists were demographically homogeneous, comprising a monoculture of knowers who mistook the uniformity of their standpoints for an objective, unbiased view from nowhere[152–155]. It was only after science became more demographically diverse that the existence of this monoculture became identifiable. At this point, scholars also came to recognize how the standpoints of that monoculture were embedded in scientific claims[156] and began distinguishing between 'strong objectivity' (which accounts for the embodied standpoints of researchers) and 'weak objectivity' (which fails to recognize the existence of standpoints at all)[157]. Strong objectivity improves scientific practice not just by recognizing the potential biasing influence of individual standpoints, but also by embracing the diversity of standpoints as a source of innovation and scientific robustness[158]. Similar to how we subject research to peer review (and thus multiple knowers), diversifying the types of knowers involved in scientific knowledge production will strengthen the emergent findings.

Decades after strong objectivity was first theorized, there is now abundant empirical evidence that cognitive and demographic diversity are important for the scientific project[124,159]. Cognitively diverse teams

Box 3

Questions for future research

Philosophical questions

- What is the nature of knowledge in predictive and generative AI? How does this knowledge differ from human knowledge?
- Could AI systems be capable of understanding, and if so, how would it differ from human understanding? What kinds of epistemic dependence arise in human–AI collaborations?
- How should theories of scientific understanding be updated as AI tools are incorporated into scientific knowledge production?

Cognitive science questions

- Are there other illusions of understanding that are unique to human–AI collaborations?
- Does expertise in a particular scientific domain protect against illusions of understanding in that domain?
- What kinds of intervention can protect against illusions of understanding in science?
- Do empirical studies of trust in AI have an expiration date? That is, as AI technologies develop and public understanding of them evolves, how often do we need to update our scientific evidence about attitudes towards AI?

AI for science questions

- How can researchers developing AI tools for scientific discovery communicate the epistemic risks of their work in a way that is accurate and accessible to non-experts?
- What incentives are there for prioritizing interpretability in ways that make the risks and benefits of AI tools transparent for scientists?
- How can conflicts of interest be navigated, particularly in cases in which there may be intellectual and/or financial incentives to oversell the capabilities of AI tools?
- Do these considerations apply differently across the visions of Oracle, Surrogate, Quant and Arbiter, as identified in this Perspective?

Social and political questions

- What kinds of power dynamics do scientific monocultures entrench?
- How can communities of knowledge cultivate and maintain cognitive and demographic diversity?
- Which scientists will benefit the most from AI productivity gains, and how will this affect existing race and gender disparities in scientific training and career advancement?
- How will reliance on AI tools affect public trust in science?

are better at problem-solving[123,160] (especially when the problems are complex[161]) and have been shown to produce patents with higher quality and impact[162]. Teams that are more demographically diverse in terms of ethnicity[163] and gender[164–166] produce more impactful science, as measured by citations. The research topics that scientists choose are correlated with their race and gender, suggesting that demographic diversity expands the coverage of topics studied[167]. But scientists do not always reap the benefits of diversity. In particular, although the contributions of scholars from minority groups tend to be more innovative, they receive less uptake from fellow scientists[122,167,168]. Thus, the practice of science still needs to catch up with the theorized benefits of having diverse knowers.

The unrealized promises of strong objectivity underscore the distinct risks posed by visions of AI and the monoculture of knowers they invite. First, there is the risk of rehomogenizing the scientific ecosystem[169], which has only recently made strides in diversifying the pool of knowers and still has considerable progress to make in terms of ensuring enduring and robust diversity[170]. Second, there is the risk of reproducing the weak objectivity of the past, failing to appreciate that AI tools embed the largely homogeneous standpoints of their creators as well as those of the dominant social groups[7,171,172]. Visions of AI for science invite an illusion of objectivity, in which scientists falsely believe that AI tools either eliminate all standpoints (in the case of Oracles and Arbiters) or are capable of representing everyone (as desired for Surrogates) (Fig. 1c). By reasserting the fantasy of a single kind of knower masked as neutral and universal (but actually reflecting the standpoints of the AI tool builders), visions of 'objective' AI tools for science retreat from recent progress in recognizing the necessity of diverse standpoints for the scientific project.

Looking ahead

Scientists must consider not only the technical limitations and potential of AI, but also how it stands to affect the social practices of scientific knowledge production (Box 2). We have analysed what makes AI tools so compelling for scientists and how these desires give rise to specific visions of AI across the research pipeline. We show how these visions foreshadow a future for science that lacks diversity, not only in terms of participants but also in terms of the research topics pursued. They also invite illusions of understanding that prevent us from appreciating how our view has narrowed. The challenge, then, is to determine how we can leverage the scientific potential of AI tools without cultivating scientific monocultures, and while remaining aware that increasing productivity does not guarantee an improved understanding of the world. We conclude with several suggestions for navigating this complex trade-off.

Any conversation about AI in science must include a reminder that AI is not a monolith. Our framework distinguishes between the visions of Oracle, Surrogate, Quant and Arbiter, and it invites researchers to be clear about why they want to use AI in their research. Doing so will help to identify which visions they might be invoking and therefore which epistemic risks might manifest in their work. There may also be cases in which risks are low, such as using AI for routine tasks (such as composing emails) or tasks within one's domain of expertise (such as developing code that one is capable of programming oneself, given enough time). Future work should explore how a researcher's expertise and stage of training affects their susceptibility to the epistemic risks of AI (Box 3). We note, however, that the risks of monocultures remain even when AI tools are being implemented by competent users, because the productivity advantages they offer can cause AI-led science, and its associated monocultures of knowing and knower, to proliferate.

One strategy to mitigate the epistemic risks that individuals might face is to work in cognitively and demographically diverse teams. For example, interdisciplinary teams produce more valid and trustworthy AI research in biology and medicine[117,173,174]. However, visions of self-driving laboratories and AI-powered research assistants suggest that AI tools are being seen as capable of eventually replacing human collaborators. Even at present, it is tempting to use an AI tool that can expand a team's domain expertise. In such settings, team members are less qualified to evaluate whether an AI model produces reasonable results[117] while simultaneously being more prone to trust AI[116]. The efficiencies gained by relying on AI tools to expand the expertise of a team must be weighed against the costs of erasing diverse, human standpoints.

Even if research teams are diversely composed, there are additional risks to consider when using AI tools created by private companies. Google and Microsoft are invested (and investing) in researchers using their AI models[175,176]. Assessing the credibility of these models is challenging because industry-developed models are often considered to be trade secrets and are thus less transparent and reproducible than models developed in academia[177,178]. Scientists must also consider how the goals of industry (which often diverge from the goals of science) shape and constrain the development of AI tools[11,177].

Although scientists who are not themselves AI researchers might be inclined to defer to AI experts as we navigate the potentials and risks of AI for science, this Perspective is itself an example of cultivating a conversation that resists the spread of a scientific monoculture. By offering a framework that is emergent from ideas in the humanities and qualitative social sciences, we hope to illustrate the benefit of drawing on diverse kinds of knowledge to strengthen scientific practices. Training the next generation of scientists to identify and avoid the epistemic risks of AI will require not only technical education, but also exposure to scholarship in science and technology studies, social epistemology and philosophy of science.

Scientific knowledge production is a fundamentally social practice that is shaped by the norms of its institutions[148]. The proposed visions of AI make it clear that a primary motivation for these tools emerges from the impulse to produce more science, more quickly and more cheaply. Given evidence that increased publications stagnate the generation of new ideas[179,180], considering the epistemic risks of AI provides us with an opportunity to reflect on whether this level of productivity—one demanded by academic and publishing institutions—is one that researchers desire and one that benefits the collective endeavour of scientific understanding[181,182]. Although visions of AI in science paint its widespread adoption as both inevitable and desirable, we should remember that scientists have a say in how things proceed. We decide when and how AI deserves to be included in our communities of knowledge.

1. Crabtree, G. Self-driving laboratories coming of age. *Joule* **4**, 2538–2541 (2020).
2. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
   This review explores how AI can be incorporated across the research pipeline, drawing from a wide range of scientific disciplines.
3. Dillion, D., Tandon, N., Gu, Y. & Gray, K. Can AI language models replace human participants? *Trends Cogn. Sci.* **27**, 597–600 (2023).
4. Grossmann, I. et al. AI and the transformation of social science research. *Science* **380**, 1108–1109 (2023).
   This forward-looking article proposes a variety of ways to incorporate generative AI into social-sciences research.
5. Gil, Y. Will AI write scientific papers in the future? *AI Mag.* **42**, 3–15 (2022).
6. Kitano, H. Nobel Turing Challenge: creating the engine for scientific discovery. *npj Syst. Biol. Appl.* **7**, 29 (2021).
7. Benjamin, R. *Race After Technology: Abolitionist Tools for the New Jim Code* (Oxford Univ. Press, 2020).
   This book examines how social norms about race become embedded in technologies, even those that are focused on providing good societal outcomes.
8. Broussard, M. *More Than a Glitch: Confronting Race, Gender, and Ability Bias in Tech* (MIT Press, 2023).
9. Noble, S. U. *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York Univ. Press, 2018).
10. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be too big? in *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623 (Association for Computing Machinery, 2021).
    One of the first comprehensive critiques of large language models, this article draws attention to a host of issues that ought to be considered before taking up such tools.
11. Crawford, K. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale Univ. Press, 2021).
12. Johnson, D. G. & Verdicchio, M. Reframing AI discourse. *Minds Mach.* **27**, 575–590 (2017).
13. Atanasoski, N. & Vora, K. *Surrogate Humanity: Race, Robots, and the Politics of Technological Futures* (Duke Univ. Press, 2019).
14. Mitchell, M. & Krakauer, D. C. The debate over understanding in AI's large language models. *Proc. Natl Acad. Sci. USA* **120**, e2215907120 (2023).
15. Kidd, C. & Birhane, A. How AI can distort human beliefs. *Science* **380**, 1222–1223 (2023).
16. Birhane, A., Kasirzadeh, A., Leslie, D. & Wachter, S. Science in the age of large language models. *Nat. Rev. Phys.* **5**, 277–280 (2023).
17. Kapoor, S. & Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **4**, 100804 (2023).
18. Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A. & Narayanan, A. The worst of both worlds: a comparative analysis of errors in learning from data in psychology and machine learning. In *Proc. 2022 AAAI/ACM Conference on AI, Ethics, and Society* (eds Conitzer, V. et al.) 335–348 (Association for Computing Machinery, 2022).
19. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
    This paper articulates the problems with attempting to explain AI systems that lack interpretability, and advocates for building interpretable models instead.
20. Crockett, M. J., Bai, X., Kapoor, S., Messeri, L. & Narayanan, A. The limitations of machine learning models for predicting scientific replicability. *Proc. Natl Acad. Sci. USA* **120**, e2307596120 (2023).
21. Lazar, S. & Nelson, A. AI safety on whose terms? *Science* **381**, 138 (2023).
22. Collingridge, D. *The Social Control of Technology* (St Martin's Press, 1980).
23. Wagner, G., Lukyanenko, R. & Paré, G. Artificial intelligence and the conduct of literature reviews. *J. Inf. Technol.* **37**, 209–226 (2022).
24. Hutson, M. Artificial-intelligence tools aim to tame the coronavirus literature. *Nature* https://doi.org/10.1038/d41586-020-01733-7 (2020).
25. Haas, Q. et al. Utilizing artificial intelligence to manage COVID-19 scientific evidence torrent with Risklick AI: a critical tool for pharmacology and therapy development. *Pharmacology* **106**, 244–253 (2021).
26. Müller, H., Pachnanda, S., Pahl, F. & Rosenqvist, C. The application of artificial intelligence on different types of literature reviews – a comparative study. In *2022 International Conference on Applied Artificial Intelligence (ICAPAI)* https://doi.org/10.1109/ICAPAI55158.2022.9801564 (Institute of Electrical and Electronics Engineers, 2022).
27. van Dinter, R., Tekinerdogan, B. & Catal, C. Automation of systematic literature reviews: a systematic literature review. *Inf. Softw. Technol.* **136**, 106589 (2021).
28. Aydın, Ö. & Karaarslan, E. OpenAI ChatGPT generated literature review: digital twin in healthcare. In *Emerging Computer Technologies 2* (ed. Aydın, Ö.) 22–31 (İzmir Akademi Dernegi, 2022).
29. AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics* **35**, 4862–4865 (2019).
30. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
31. Lee, J. S., Kim, J. & Kim, P. M. Score-based generative modeling for de novo protein design. *Nat. Computat. Sci.* **3**, 382–392 (2023).
32. Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).
33. Krenn, M. et al. On scientific understanding with artificial intelligence. *Nat. Rev. Phys.* **4**, 761–769 (2022).
34. Extance, A. How AI technology can tame the scientific literature. *Nature* **561**, 273–274 (2018).
35. Hastings, J. *AI for Scientific Discovery* (CRC Press, 2023).
    This book reviews current and future incorporation of AI into the scientific research pipeline.
36. Ahmed, A. et al. The future of academic publishing. *Nat. Hum. Behav.* **7**, 1021–1026 (2023).
37. Gray, K., Yam, K. C., Zhen'An, A. E., Wilbanks, D. & Waytz, A. The psychology of robots and artificial intelligence. In *The Handbook of Social Psychology* (eds Gilbert, D. et al.) (in the press).
38. Argyle, L. P. et al. Out of one, many: using language models to simulate human samples. *Polit. Anal.* **31**, 337–351 (2023).
39. Aher, G., Arriaga, R. I. & Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In *Proc. 40th International Conference on Machine Learning* (eds Krause, A. et al.) 337–371 (JMLR.org, 2023).
40. Binz, M. & Schulz, E. Using cognitive psychology to understand GPT-3. *Proc. Natl Acad. Sci. USA* **120**, e2218523120 (2023).
41. Ornstein, J. T., Blasingame, E. N. & Truscott, J. S. How to train your stochastic parrot: large language models for political texts. *Github*, https://joeornstein.github.io/publications/ornstein-blasingame-truscott.pdf (2023).
42. He, S. et al. Learning to predict the cosmological structure formation. *Proc. Natl Acad. Sci. USA* **116**, 13825–13832 (2019).
43. Mahmood, F. et al. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Trans. Med. Imaging* **39**, 3257–3267 (2020).
44. Teixeira, B. et al. Generating synthetic X-ray images of a person from the surface geometry. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 9059–9067 (Institute of Electrical and Electronics Engineers, 2018).
45. Marouf, M. et al. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nat. Commun.* **11**, 166 (2020).
46. Watts, D. J. A twenty-first century science. *Nature* **445**, 489 (2007).
47. boyd, d. & Crawford, K. Critical questions for big data. *Inf. Commun. Soc.* **15**, 662–679 (2012).
    This article assesses the ethical and epistemic implications of scientific and societal moves towards big data and provides a parallel case study for thinking about the risks of artificial intelligence.
48. Jolly, E. & Chang, L. J. The Flatland fallacy: moving beyond low–dimensional thinking. *Top. Cogn. Sci.* **11**, 433–454 (2019).
49. Yarkoni, T. & Westfall, J. Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* **12**, 1100–1122 (2017).
50. Radivojac, P. et al. A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10**, 221–227 (2013).
51. Bileschi, M. L. et al. Using deep learning to annotate the protein universe. *Nat. Biotechnol.* **40**, 932–937 (2022).
52. Barkas, N. et al. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods* **16**, 695–698 (2019).
53. Demszky, D. et al. Using large language models in psychology. *Nat. Rev. Psychol.* **2**, 688–701 (2023).
54. Karjus, A. Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence. Preprint at https://arxiv.org/abs/2309.14379 (2023).

55. Davies, A. et al. Advancing mathematics by guiding human intuition with AI. *Nature* **600**, 70–74 (2021).
56. Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D. & Griffiths, T. L. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science* **372**, 1209–1214 (2021).
57. Ilyas, A. et al. Adversarial examples are not bugs, they are features. Preprint at https://doi.org/10.48550/arXiv.1905.02175 (2019)
58. Semel, B. M. Listening like a computer: attentional tensions and mechanized care in psychiatric digital phenotyping. *Sci. Technol. Hum. Values* **47**, 266–290 (2022).
59. Gil, Y. Thoughtful artificial intelligence: forging a new partnership for data science and scientific discovery. *Data Sci.* **1**, 119–129 (2017).
60. Checco, A., Bracciale, L., Loreti, P., Pinfield, S. & Bianchi, G. AI-assisted peer review. *Humanit. Soc. Sci. Commun.* **8**, 25 (2021).
61. Thelwall, M. Can the quality of published academic journal articles be assessed with machine learning? *Quant. Sci. Stud.* **3**, 208–226 (2022).
62. Dhar, P. Peer review of scholarly research gets an AI boost. *IEEE Spectrum* spectrum.ieee.org/peer-review-of-scholarly-research-gets-an-ai-boost (2020).
63. Heaven, D. AI peer reviewers unleashed to ease publishing grind. *Nature* **563**, 609–610 (2018).
64. Conroy, G. How ChatGPT and other AI tools could disrupt scientific publishing. *Nature* **622**, 234–236 (2023).
65. Nosek, B. A. et al. Replicability, robustness, and reproducibility in psychological science. *Annu. Rev. Psychol.* **73**, 719–748 (2022).
66. Altmejd, A. et al. Predicting the replicability of social science lab experiments. *PLoS ONE* **14**, e0225826 (2019).
67. Yang, Y., Youyou, W. & Uzzi, B. Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proc. Natl Acad. Sci. USA* **117**, 10762–10768 (2020).
68. Youyou, W., Yang, Y. & Uzzi, B. A discipline-wide investigation of the replicability of psychology papers over the past two decades. *Proc. Natl Acad. Sci. USA* **120**, e2208863120 (2023).
69. Rabb, N., Fernbach, P. M. & Sloman, S. A. Individual representation in a community of knowledge. *Trends Cogn. Sci.* **23**, 891–902 (2019).
This comprehensive review paper documents the empirical evidence for distributed cognition in communities of knowledge and the resultant vulnerabilities to illusions of understanding.
70. Rozenblit, L. & Keil, F. The misunderstood limits of folk science: an illusion of explanatory depth. *Cogn. Sci.* **26**, 521–562 (2002).
This paper provided an empirical demonstration of the illusion of explanatory depth, and inspired a programme of research in cognitive science on communities of knowledge.
71. Hutchins, E. *Cognition in the Wild* (MIT Press, 1995).
72. Lave, J. & Wenger, E. *Situated Learning: Legitimate Peripheral Participation* (Cambridge Univ. Press, 1991).
73. Kitcher, P. The division of cognitive labor. *J. Philos.* **87**, 5–22 (1990).
74. Hardwig, J. Epistemic dependence. *J. Philos.* **82**, 335–349 (1985).
75. Keil, F. in *Oxford Studies In Epistemology* (eds Gendler, T. S. & Hawthorne, J.) 143–166 (Oxford Academic, 2005).
76. Weisberg, M. & Muldoon, R. Epistemic landscapes and the division of cognitive labor. *Philos. Sci.* **76**, 225–252 (2009).
77. Sloman, S. A. & Rabb, N. Your understanding is my understanding: evidence for a community of knowledge. *Psychol. Sci.* **27**, 1451–1460 (2016).
78. Wilson, R. A. & Keil, F. The shadows and shallows of explanation. *Minds Mach.* **8**, 137–159 (1998).
79. Keil, F. C., Stein, C., Webb, L., Billings, V. D. & Rozenblit, L. Discerning the division of cognitive labor: an emerging understanding of how knowledge is clustered in other minds. *Cogn. Sci.* **32**, 259–300 (2008).
80. Sperber, D. et al. Epistemic vigilance. *Mind Lang.* **25**, 359–393 (2010).
81. Wilkenfeld, D. A., Plunkett, D. & Lombrozo, T. Depth and deference: when and why we attribute understanding. *Philos. Stud.* **173**, 373–393 (2016).
82. Sparrow, B., Liu, J. & Wegner, D. M. Google effects on memory: cognitive consequences of having information at our fingertips. *Science* **333**, 776–778 (2011).
83. Fisher, M., Goddu, M. K. & Keil, F. C. Searching for explanations: how the internet inflates estimates of internal knowledge. *J. Exp. Psychol. Gen.* **144**, 674–687 (2015).
84. De Freitas, J., Agarwal, S., Schmitt, B. & Haslam, N. Psychological factors underlying attitudes toward AI tools. *Nat. Hum. Behav.* **7**, 1845–1854 (2023).
85. Castelo, N., Bos, M. W. & Lehmann, D. R. Task-dependent algorithm aversion. *J. Mark. Res.* **56**, 809–825 (2019).
86. Cadario, R., Longoni, C. & Morewedge, C. K. Understanding, explaining, and utilizing medical artificial intelligence. *Nat. Hum. Behav.* **5**, 1636–1642 (2021).
87. Oktar, K. & Lombrozo, T. Deciding to be authentic: intuition is favored over deliberation when authenticity matters. *Cognition* **223**, 105021 (2022).
88. Bigman, Y. E., Yam, K. C., Marciano, D., Reynolds, S. J. & Gray, K. Threat of racial and economic inequality increases preference for algorithm decision-making. *Comput. Hum. Behav.* **122**, 106859 (2021).
89. Claudy, M. C., Aquino, K. & Graso, M. Artificial intelligence can't be charmed: the effects of impartiality on laypeople's algorithmic preferences. *Front. Psychol.* **13**, 898027 (2022).
90. Snyder, C., Keppler, S. & Leider, S. Algorithm reliance under pressure: the effect of customer load on service workers. Preprint at *SSRN* https://doi.org/10.2139/ssrn.4066823 (2022).
91. Bogert, E., Schecter, A. & Watson, R. T. Humans rely more on algorithms than social influence as a task becomes more difficult. *Sci Rep.* **11**, 8028 (2021).
92. Raviv, A., Bar-Tal, D., Raviv, A. & Abin, R. Measuring epistemic authority: studies of politicians and professors. *Eur. J. Personal.* **7**, 119–138 (1993).
93. Cummings, L. The "trust" heuristic: arguments from authority in public health. *Health Commun.* **29**, 1043–1056 (2014).
94. Lee, M. K. Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management. *Big Data Soc.* **5**, https://doi.org/10.1177/2053951718756684 (2018).
95. Kissinger, H. A., Schmidt, E. & Huttenlocher, D. *The Age of A.I. And Our Human Future* (Little, Brown, 2021).
96. Lombrozo, T. Explanatory preferences shape learning and inference. *Trends Cogn. Sci.* **20**, 748–759 (2016).
This paper provides an overview of philosophical theories of explanatory virtues and reviews empirical evidence on the sorts of explanations people find satisfying.
97. Vrantsidis, T. H. & Lombrozo, T. Simplicity as a cue to probability: multiple roles for simplicity in evaluating explanations. *Cogn. Sci.* **46**, e13169 (2022).
98. Johnson, S. G. B., Johnston, A. M., Toig, A. E. & Keil, F. C. Explanatory scope informs causal strength inferences. In *Proc. 36th Annual Meeting of the Cognitive Science Society* 2453–2458 (Cognitive Science Society, 2014).
99. Khemlani, S. S., Sussman, A. B. & Oppenheimer, D. M. Harry Potter and the sorcerer's scope: latent scope biases in explanatory reasoning. *Mem. Cognit.* **39**, 527–535 (2011).
100. Liquin, E. G. & Lombrozo, T. Motivated to learn: an account of explanatory satisfaction. *Cogn. Psychol.* **132**, 101453 (2022).
101. Hopkins, E. J., Weisberg, D. S. & Taylor, J. C. V. The seductive allure is a reductive allure: people prefer scientific explanations that contain logically irrelevant reductive information. *Cognition* **155**, 67–76 (2016).
102. Weisberg, D. S., Hopkins, E. J. & Taylor, J. C. V. People's explanatory preferences for scientific phenomena. *Cogn. Res. Princ. Implic.* **3**, 44 (2018).
103. Jerez-Fernandez, A., Angulo, A. N. & Oppenheimer, D. M. Show me the numbers: precision as a cue to others' confidence. *Psychol. Sci.* **25**, 633–635 (2014).
104. Kim, J., Giroux, M. & Lee, J. C. When do you trust AI? The effect of number presentation detail on consumer trust and acceptance of AI recommendations. *Psychol. Mark.* **38**, 1140–1155 (2021).
105. Nguyen, C. T. The seductions of clarity. *R. Inst. Philos. Suppl.* **89**, 227–255 (2021).
This article describes how reductive and quantitative explanations can generate a sense of understanding that is not necessarily correlated with actual understanding.
106. Fisher, M., Smiley, A. H. & Grillo, T. L. H. Information without knowledge: the effects of internet search on learning. *Memory* **30**, 375–387 (2022).
107. Eliseev, E. D. & Marsh, E. J. Understanding why searching the internet inflates confidence in explanatory ability. *Appl. Cogn. Psychol.* **37**, 711–720 (2023).
108. Fisher, M. & Oppenheimer, D. M. Who knows what? Knowledge misattribution in the division of cognitive labor. *J. Exp. Psychol. Appl.* **27**, 292–306 (2021).
109. Chromik, M., Eiband, M., Buchner, F., Krüger, A. & Butz, A. I think I get your point, AI! The illusion of explanatory depth in explainable AI. In *26th International Conference on Intelligent User Interfaces* (eds Hammond, T. et al.) 307–317 (Association for Computing Machinery, 2021).
110. Strevens, M. No understanding without explanation. *Stud. Hist. Philos. Sci. A* **44**, 510–515 (2013).
111. Ylikoski, P. in *Scientific Understanding: Philosophical Perspectives* (eds De Regt, H. et al.) 100–119 (Univ. Pittsburgh Press, 2009).
112. Giudice, M. D. The prediction–explanation fallacy: a pervasive problem in scientific applications of machine learning. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/4vq8f (2021).
113. Hofman, J. M. et al. Integrating explanation and prediction in computational social science. *Nature* **595**, 181–188 (2021).
This paper highlights the advantages and disadvantages of explanatory versus predictive approaches to modelling, with a focus on applications to computational social science.
114. Shmueli, G. To explain or to predict? *Stat. Sci.* **25**, 289–310 (2010).
115. Hofman, J. M., Sharma, A. & Watts, D. J. Prediction and explanation in social systems. *Science* **355**, 486–488 (2017).
116. Logg, J. M., Minson, J. A. & Moore, D. A. Algorithm appreciation: people prefer algorithmic to human judgment. *Organ. Behav. Hum. Decis. Process.* **151**, 90–103 (2019).
117. Nguyen, C. T. Cognitive islands and runaway echo chambers: problems for epistemic dependence on experts. *Synthese* **197**, 2803–2821 (2020).
118. Breiman, L. Statistical modeling: the two cultures. *Stat. Sci.* **16**, 199–215 (2001).
119. Gao, J. & Wang, D. Quantifying the benefit of artificial intelligence for scientific research. Preprint at arxiv.org/abs/2304.10578 (2023).
120. Hanson, B. et al. Garbage in, garbage out: mitigating risks and maximizing benefits of AI in research. *Nature* **623**, 28–31 (2023).
121. Kleinberg, J. & Raghavan, M. Algorithmic monoculture and social welfare. *Proc. Natl Acad. Sci. USA* **118**, e2018340118 (2021).
This paper uses formal modelling methods to demonstrate that when companies all rely on the same algorithm to make decisions (an algorithmic monoculture), the overall quality of those decisions is reduced because valuable options can slip through the cracks, even when the algorithm performs accurately for individual companies.
122. Hofstra, B. et al. The diversity–innovation paradox in science. *Proc. Natl Acad. Sci. USA* **117**, 9284–9291 (2020).
123. Hong, L. & Page, S. E. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl Acad. Sci. USA* **101**, 16385–16389 (2004).
124. Page, S. E. Where diversity comes from and why it matters? *Eur. J. Soc. Psychol.* **44**, 267–279 (2014).
This article reviews research demonstrating the benefits of cognitive diversity and diversity in methodological approaches for problem solving and innovation.
125. Clarke, A. E. & Fujimura, J. H. (eds) *The Right Tools for the Job: At Work in Twentieth-Century Life Sciences* (Princeton Univ. Press, 2014).
126. Silva, V. J., Bonacelli, M. B. M. & Pacheco, C. A. Framing the effects of machine learning on science. *AI Soc.* https://doi.org/10.1007/s00146-022-01515-x (2022).
127. Sassenberg, K. & Ditrich, L. Research in social psychology changed between 2011 and 2016: larger sample sizes, more self-report measures, and more online studies. *Adv. Methods Pract. Psychol. Sci.* **2**, 107–114 (2019).
128. Simon, A. F. & Wilder, D. Methods and measures in social and personality psychology: a comparison of JPSP publications in 1982 and 2016. *J. Soc. Psychol.* https://doi.org/10.1080/00224545.2022.2135088 (2022).

129. Anderson, C. A. et al. The MTurkification of social and personality psychology. *Pers. Soc. Psychol. Bull.* **45**, 842–850 (2019).
130. Latour, B. in *The Social After Gabriel Tarde: Debates and Assessments* (ed. Candea, M.) 145–162 (Routledge, 2010).
131. Porter, T. M. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton Univ. Press, 1996).
132. Lazer, D. et al. Meaningful measures of human society in the twenty-first century. *Nature* **595**, 189–196 (2021).
133. Knox, D., Lucas, C. & Cho, W. K. T. Testing causal theories with learned proxies. *Annu. Rev. Polit. Sci.* **25**, 419–441 (2022).
134. Barberá, P. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Polit. Anal.* **23**, 76–91 (2015).
135. Brady, W. J., McLoughlin, K., Doan, T. N. & Crockett, M. J. How social learning amplifies moral outrage expression in online social networks. *Sci. Adv.* **7**, eabe5641 (2021).
136. Barnes, J., Klinger, R. & im Walde, S. S. Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets. In *Proc. 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (eds Balahur, A. et al.) 2–12 (Association for Computational Linguistics, 2017).
137. Gitelman, L. (ed.) *"Raw Data" is an Oxymoron* (MIT Press, 2013).
138. Breznau, N. et al. Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proc. Natl Acad. Sci. USA* **119**, e2203150119 (2022).
This study demonstrates how 73 research teams analysing the same dataset reached different conclusions about the relationship between immigration and public support for social policies, highlighting the subjectivity and uncertainty involved in analysing complex datasets.
139. Gillespie, T. in *Media Technologies: Essays on Communication, Materiality, and Society* (eds Gillespie, T. et al.) 167–194 (MIT Press, 2014).
140. Leonelli, S. *Data-Centric Biology: A Philosophical Study* (Univ. Chicago Press, 2016).
141. Wang, A., Kapoor, S., Barocas, S. & Narayanan, A. Against predictive optimization: on the legitimacy of decision-making algorithms that optimize predictive accuracy. *ACM J. Responsib. Comput.*, https://doi.org/10.1145/3636509 (2023).
142. Athey, S. Beyond prediction: using big data for policy problems. *Science* **355**, 483–485 (2017).
143. del Rosario Martínez-Ordaz, R. Scientific understanding through big data: from ignorance to insights to understanding. *Possibility Stud. Soc.* **1**, 279–299 (2023).
144. Nussberger, A.-M., Luo, L., Celis, L. E. & Crockett, M. J. Public attitudes value interpretability but prioritize accuracy in artificial intelligence. *Nat. Commun.* **13**, 5821 (2022).
145. Zittrain, J. in *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives* (eds. Voeneky, S. et al.) 176–184 (Cambridge Univ. Press, 2022).
This article articulates the epistemic risks of prioritizing predictive accuracy over explanatory understanding when AI tools are interacting in complex systems.
146. Shumailov, I. et al. The curse of recursion: training on generated data makes models forget. Preprint at arxiv.org/abs/2305.17493 (2023).
147. Latour, B. *Science In Action: How to Follow Scientists and Engineers Through Society* (Harvard Univ. Press, 1987).
This book provides strategies and approaches for thinking about science as a social endeavour.
148. Franklin, S. Science as culture, cultures of science. *Annu. Rev. Anthropol.* **24**, 163–184 (1995).
149. Haraway, D. Situated knowledges: the science question in feminism and the privilege of partial perspective. *Fem. Stud.* **14**, 575–599 (1988).
This article acknowledges that the objective 'view from nowhere' is unobtainable: knowledge, it argues, is always situated.
150. Harding, S. *Objectivity and Diversity: Another Logic of Scientific Research* (Univ. Chicago Press, 2015).
151. Longino, H. E. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry* (Princeton Univ. Press, 1990).
152. Daston, L. & Galison, P. *Objectivity* (Princeton Univ. Press, 2007).
This book is a historical analysis of the shifting modes of 'objectivity' that scientists have pursued, arguing that objectivity is not a universal concept but that it shifts alongside scientific techniques and ambitions.
153. Prescod-Weinstein, C. Making Black women scientists under white empiricism: the racialization of epistemology in physics. *Signs J. Women Cult. Soc.* **45**, 421–447 (2020).
154. Mavhunga, C. *What Do Science, Technology, and Innovation Mean From Africa?* (MIT Press, 2017).
155. Schiebinger, L. *The Mind Has No Sex? Women in the Origins of Modern Science* (Harvard Univ. Press, 1991).
156. Martin, E. The egg and the sperm: how science has constructed a romance based on stereotypical male–female roles. *Signs J. Women Cult. Soc.* **16**, 485–501 (1991).
This case study shows how assumptions about gender affect scientific theories, sometimes delaying the articulation of what might be considered to be more accurate descriptions of scientific phenomena.
157. Harding, S. Rethinking standpoint epistemology: What is "strong objectivity"? *Centen. Rev.* **36**, 437–470 (1992).
In this article, Harding outlines her position on 'strong objectivity', by which clearly articulating one's standpoint can lead to more robust knowledge claims.
158. Oreskes, N. *Why Trust Science?* (Princeton Univ. Press, 2019).
This book introduces the reader to 20 years of scholarship in science and technology studies, arguing that the tools the discipline has for understanding science can help to reinstate public trust in the institution.
159. Rolin, K., Koskinen, I., Kuorikoski, J. & Reijula, S. Social and cognitive diversity in science: introduction. *Synthese* **202**, 36 (2023).
160. Hong, L. & Page, S. E. Problem solving by heterogeneous agents. *J. Econ. Theory* **97**, 123–163 (2001).
161. Sulik, J., Bahrami, B. & Deroy, O. The diversity gap: when diversity matters for knowledge. *Perspect. Psychol. Sci.* **17**, 752–767 (2022).
162. Lungeanu, A., Whalen, R., Wu, Y. J., DeChurch, L. A. & Contractor, N. S. Diversity, networks, and innovation: a text analytic approach to measuring expertise diversity. *Netw. Sci.* **11**, 36–64 (2023).
163. AlShebli, B. K., Rahwan, T. & Woon, W. L. The preeminence of ethnic diversity in scientific collaboration. *Nat. Commun.* **9**, 5163 (2018).
164. Campbell, L. G., Mehtani, S., Dozier, M. E. & Rinehart, J. Gender-heterogeneous working groups produce higher quality science. *PLoS ONE* **8**, e79147 (2013).
165. Nielsen, M. W., Bloch, C. W. & Schiebinger, L. Making gender diversity work for scientific discovery and innovation. *Nat. Hum. Behav.* **2**, 726–734 (2018).
166. Yang, Y., Tian, T. Y., Woodruff, T. K., Jones, B. F. & Uzzi, B. Gender-diverse teams produce more novel and higher-impact scientific ideas. *Proc. Natl Acad. Sci. USA* **119**, e2200841119 (2022).
167. Kozlowski, D., Larivière, V., Sugimoto, C. R. & Monroe-White, T. Intersectional inequalities in science. *Proc. Natl Acad. Sci. USA* **119**, e2113067119 (2022).
168. Fehr, C. & Jones, J. M. Culture, exploitation, and epistemic approaches to diversity. *Synthese* **200**, 465 (2022).
169. Nakadai, R., Nakawake, Y. & Shibasaki, S. AI language tools risk scientific diversity and innovation. *Nat. Hum. Behav.* **7**, 1804–1805 (2023).
170. National Academies of Sciences, Engineering, and Medicine et al. *Advancing Antiracism, Diversity, Equity, and Inclusion in STEMM Organizations: Beyond Broadening Participation* (National Academies Press, 2023).
171. Winner, L. Do artifacts have politics? *Daedalus* **109**, 121–136 (1980).
172. Eubanks, V. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St. Martin's Press, 2018).
173. Littmann, M. et al. Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. *Nat. Mach. Intell.* **2**, 18–24 (2020).
174. Carusi, A. et al. Medical artificial intelligence is as much social as it is technological. *Nat. Mach. Intell.* **5**, 98–100 (2023).
175. Raghu, M. & Schmidt, E. A survey of deep learning for scientific discovery. Preprint at arxiv.org/abs/2003.11755 (2020).
176. Bishop, C. AI4Science to empower the fifth paradigm of scientific discovery. *Microsoft Research Blog* www.microsoft.com/en-us/research/blog/ai4science-to-empower-the-fifth-paradigm-of-scientific-discovery/ (2022).
177. Whittaker, M. The steep cost of capture. *Interactions* **28**, 50–55 (2021).
178. Liesenfeld, A., Lopez, A. & Dingemanse, M. Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proc. 5th International Conference on Conversational User Interfaces* 1–6 (Association for Computing Machinery, 2023).
179. Chu, J. S. G. & Evans, J. A. Slowed canonical progress in large fields of science. *Proc. Natl Acad. Sci. USA* **118**, e2021636118 (2021).
180. Park, M., Leahey, E. & Funk, R. J. Papers and patents are becoming less disruptive over time. *Nature* **613**, 138–144 (2023).
181. Frith, U. Fast lane to slow science. *Trends Cogn. Sci.* **24**, 1–2 (2020).
This article explains the epistemic risks of a hyperfocus on scientific productivity and explores possible avenues for incentivizing the production of higher-quality science on a slower timescale.
182. Stengers, I. *Another Science is Possible: A Manifesto for Slow Science* (Wiley, 2018).
183. Lake, B. M. & Baroni, M. Human-like systematic generalization through a meta-learning neural network. *Nature* **623**, 115–121 (2023).
184. Feinman, R. & Lake, B. M. Learning task-general representations with generative neuro-symbolic modeling. Preprint at arxiv.org/abs/2006.14448 (2021).
185. Schölkopf, B. et al. Toward causal representation learning. *Proc. IEEE* **109**, 612–634 (2021).
186. Mitchell, M. AI's challenge of understanding the world. *Science* **382**, eadm8175 (2023).
187. Sartori, L. & Bocca, G. Minding the gap(s): public perceptions of AI and socio-technical imaginaries. *AI Soc.* **38**, 443–458 (2023).

Acknowledgements We thank D. S. Bassett, W. J. Brady, S. Helmreich, S. Kapoor, T. Lombrozo, A. Narayanan, M. Salganik and A. J. te Velthuis for comments. We also thank C. Buckner and P. Winter for their feedback and suggestions.

Author contributions The authors contributed equally to the research and writing of the paper.

Competing interests The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Lisa Messeri or M. J. Crockett.

Peer review information *Nature* thanks Cameron Buckner, Peter Winter and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature Limited 2024