PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Opinion piece

Cite this article: Gweon H, Fan J, Kim B. 2023 Socially intelligent machines that learn from humans and help humans learn. *Phil. Trans. R. Soc. A* **381**: 20220048.
https://doi.org/10.1098/rsta.2022.0048

Received: 19 December 2022
Accepted: 17 April 2023

One contribution of 11 to a discussion meeting issue 'Cognitive artificial intelligence'.

Subject Areas:
artificial intelligence

Keywords:
artificial intelligence, social intelligence, cognitive science, theory of mind, communication

Author for correspondence:
Hyowon Gweon
e-mail: gweon@stanford.edu

Socially intelligent machines that learn from humans and help humans learn

Hyowon Gweon[1], Judith Fan[1,2] and Been Kim[3]

[1]Department of Psychology, Stanford University, Stanford, CA 94305, USA
[2]Department of Psychology, University of California, San Diego, CA 92093, USA
[3]Google Research, Mountain View, CA 94043, USA

HG, 0000-0001-8012-3526; JF, 0000-0002-0097-3254

A hallmark of human intelligence is the ability to understand and influence other minds. Humans engage in inferential social learning (ISL) by using commonsense psychology to learn from others and help others learn. Recent advances in artificial intelligence (AI) are raising new questions about the feasibility of human–machine interactions that support such powerful modes of social learning. Here, we envision what it means to develop socially intelligent machines that can learn, teach, and communicate in ways that are characteristic of ISL. Rather than machines that simply predict human behaviours or recapitulate superficial aspects of human sociality (e.g. smiling, imitating), we should aim to build machines that can learn from human inputs and generate outputs for humans by proactively considering human values, intentions and beliefs. While such machines can inspire next-generation AI systems that learn more effectively from humans (as learners) and even help humans acquire new knowledge (as teachers), achieving these goals will also require scientific studies of its counterpart: how humans reason about machine minds and behaviours. We close by discussing the need for closer collaborations between the AI/ML and cognitive science communities to advance a science of both natural and artificial intelligence.

This article is part of a discussion meeting issue 'Cognitive artificial intelligence'.

THE ROYAL SOCIETY PUBLISHING

royalsocietypublishing.org/journal/rsta Phil. Trans. R. Soc. A 381: 20220048

1. Introduction

Throughout history, humans have leveraged their intelligence to accumulate shared knowledge, invent new technologies, and transform their environment to survive and flourish [1,2]. Now, driven by remarkable progress in machine learning and artificial intelligence (AI), an ever-increasing number of AI systems are influencing the way things are done, from everyday tasks (e.g. predictive search, photo tagging) to highly specialized ones (e.g. antibody discovery, nuclear fusion) [3]. For better or worse, human civilization has long benefited from finding new ways of using other species as resources and helpers. Similarly, it is now conceivable that future human endeavours will be increasingly intertwined with AI systems that assist and collaborate with humans in ever richer (and perhaps riskier) ways. Yet the challenge we face today is qualitatively different from taming existing species; we are *developing* new species—artificially engineered machines—that exhibit intelligent behaviours.

Today's AI systems often use neural networks inspired by biological brains, and even exhibit behavioural and neural signatures of perceptual and linguistic processing that resemble those of humans [4–8]. These systems, however, are far from duplicates of human intelligence; they learn, represent, and reason in ways that are qualitatively different from humans. Machines that generate text outputs with human-level fluency (e.g. ChatGPT [9]) still make mistakes that humans would not [10–12], and those that show superhuman performance in solving hard problems [13,14]) might generate solutions that are incomprehensible to humans. In order for these different species to assist and collaborate with humans effectively, it is crucial for them to understand and interpret human inputs in reasonable ways, and generate outputs that can be understood and used by humans.

The field of AI interpretability has responded to these challenges by developing various ways to make machines more 'understandable' to humans [15–19], such as exposing which features of the input were responsible for a machine's decision. For instance, an image classification decision might be 'explained' by highlighting the pixels that contributed most to a vision algorithm's prediction [20]. Despite some progress, limitations of such approaches have also become evident (e.g. [21–28]). One issue is that these tools treat the challenge of creating a communication channel from machines to humans as a pure engineering problem, aiming to achieve technical success (e.g. implementing a simple, one-way transmission mechanism that converts what a machine 'sees' into what a human sees) without facing the fundamental scientific challenge: bridging the representational gap between humans and machines to achieve mutual understanding [29].

Addressing this scientific challenge will require more than just technical innovation; we also need a framework that grounds our vision for what it means to build socially intelligent machines. Here, we propose such a framework by drawing insights from how humans achieve mutual understanding through inferential social learning (ISL, [30]). Although this idea was originally proposed as an account of how humans learn from others and help others learn, it can also serve as a useful guide for generating new research directions beyond interpretability to make human–machine interactions more informative, productive, and beneficial for humans.

Calls for AI research to take inspiration from human cognitive development are not new. For instance, the idea that human children can learn-to-learn based on abstract, causal theories about the world has been widely discussed in the context of building more human-like intelligent machines [31]. Yet, the profoundly social nature of human learning and the importance of teaching in facilitating such learning has been relatively underappreciated in the machine learning/AI community. Instead, the word 'social' has often been associated with a specific subset of research endeavours such as building 'social robots' that aim to emulate the human ability to express emotions, empathize, and connect [32,33], or 'social AI' specialized for tasks that require coordination, cooperation or negotiation (e.g. DeepNash [34], CICERO [35]). Moreover, there are a range of efforts in machine learning/AI to implement specific aspects of social learning and teaching, such as imitation learning [36,37], machine teaching [38], reinforcement learning from human feedback [39], and cooperative inverse reinforcement learning [40]. Thus there is a

royalsocietypublishing.org/journal/rsta Phil. Trans. R. Soc. A 381: 20220048

growing need for a theoretical framework that identifies the core characteristics of human-like social intelligence and contexualizes these various research efforts.

To this end, our proposal highlights the importance of social cognition: the ability to *think* about the minds of other agents. Often referred to as Theory of Mind, or commonsense psychology more broadly,[1] this ability supports both the production and interpretation of human social behaviours (e.g. gaze, point, verbal communication) and strategies (e.g. cooperate, compete, negotiate). Engaging in inferential social learning means using commonsense psychology to learn from others about the world (social learning) and share one's own knowledge with others (teaching). As such, studying inferential social learning in humans not only offers a window into how humans achieve mutual understanding, but also an opportunity to gain insights for achieving human–machine understanding.

(a) From human–human understanding to human–machine understanding

In the past decade, productive collaborations between developmental psychology and computational cognitive science have made significant advances in understanding the human mind. Examples include characterizing human learning and exploration as a process of theory-building [42–45], formalizing social cognition as inverse planning based on causal models of other agents [41,46,47], and building computational models of teaching as cooperative communication that critically relies on social cognition [48–50]. These formal accounts have been used to generate predictions for behavioural experiments with humans, especially young children [51–54]. The inferential social learning (ISL) framework—the idea that humans learn by drawing inferences from data provided by others—offers a comprehensive account of social learning, teaching and communication that synthesizes these advances [30]. Rather than focusing on a learner who explores the world or imitates another agent, or a teacher who demonstrates or instructs, it aims to *characterize the interactions between two agents—one who wants to learn (i.e. learner) and the other who wants to teach (i.e. teacher)—as the result of mutual mental-state reasoning and planning*.

While social learning has long been associated with copying, imitation, emulation, or cue-following, ISL offers a different perspective that focuses on representations of other minds and inferential processes that are powerful yet flexible and context-sensitive. Although young children tend to imitate what others do [55,56] and trust what others say [57], they are far from indiscriminate copycats or sponges; whether they imitate, trust, or follow advice depends on the social context, meaning that these behaviours are modulated by *who* is doing *what* and *why* [58–60]. ISL explains such selective, 'smart' social learning as the result of sophisticated inferences, rather than learned rules or heuristics for whom to copy and trust. It also integrates theoretical accounts of social learning and teaching into a single framework, highlighting the importance of social cognition (i.e. theory of mind, commonsense psychology) as a key prerequisite for both powerful social learning and effective teaching.

While ISL was developed as an account of human-to-human learning and teaching, we believe that situating humans and AI agents within this framework can provide a richer characterization of human–machine understanding and generate new research directions. As human interaction is premised on a shared understanding of each others' mental states, we believe that human–machine interactions should also be grounded on a shared understanding of each other. Note that we are not necessarily envisioning AI agents that fully replicate human-like social learning. After all, communication between humans is far from perfect [61], and in many contexts, it may not be the best model for human–machine intelligibility. Yet, human social learning and teaching is by far the best known example of high-bandwidth interaction between complex systems, making it an obvious starting point for imagining machines that can interact with humans as effective helpers and collaborators. Thus it is critical to identify the deeper principles that give rise to human-like

[1]Here, we use the term commonsense psychology to refer to the ability to reason about others' actions in terms of their underlying mental states (i.e. theory of mind) as well as their expected costs and rewards (i.e. Naive Utility Calculus, [41]).

royalsocietypublishing.org/journal/rsta Phil. Trans. R. Soc. A 381: 20220048

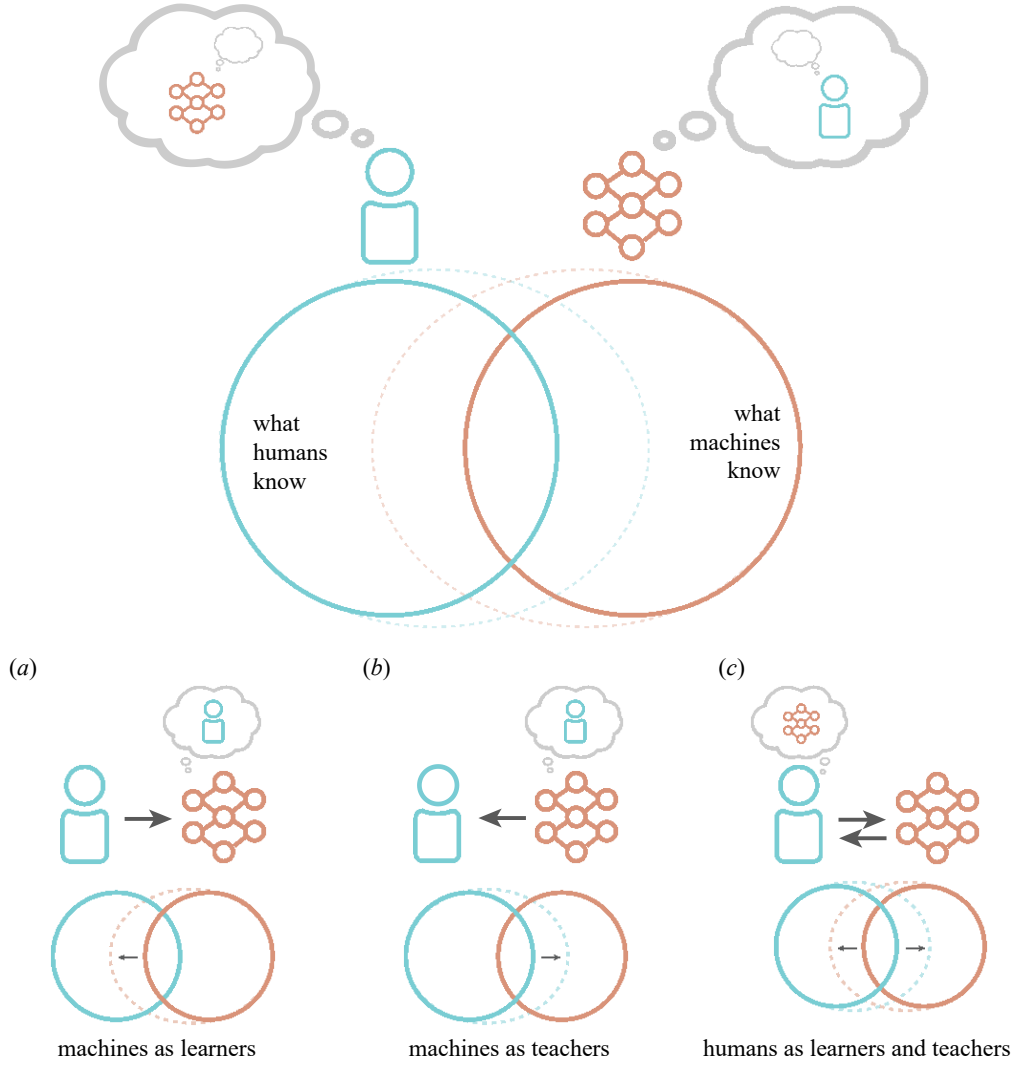

Figure 1. Inferential social learning (ISL) between humans and socially intelligent machines. Top: just as humans learn from others and help others learn by using a mental model of each others' minds (i.e. a generative model of how an agent thinks, plans and acts), humans and machines can engage in more powerful and flexible social learning and teaching by using a mental model of each other. (*a*) Socially intelligent machines would be able to learn by drawing rich inferences about human intentions from observations of their behaviour (§3(a)). (*b*) Socially intelligent machines would be able to communicate what they know to help humans learn (§3(b)). (*c*) Developing socially intelligent machines will also require understanding how humans think about machine minds as they learn from, evaluate, and teach machines (§4).

social interaction, while cautioning against approaches that merely mimic the superficial aspects of human social intelligence.

In the next section, we draw out the key aspects of ISL in humans as learners and as teachers (§2). We then apply this idea to machines and humans (figure 1) to discuss how these ideas might apply to machines as learners (§3(a), figure 1*a*) and as teachers (§3(b), figure 1*b*) that can even open up opportunities for humans to learn and benefit from new knowledge they might discover. We then argue for the need for research on how humans represent and reason about AI to achieve successful human–machine ISL (§4, figure 1*c*) and close by arguing that further progress is likely to require close coordination and collaborations between different disciplines (§5).

royalsocietypublishing.org/journal/rsta Phil. Trans. R. Soc. A 381: 20220048

2. Inferential social learning in humans

Inferential social learning (ISL) involves two agents who have distinct but intertwined goals: a teacher and a learner. Here, we use the terms 'learners' and 'teachers' more broadly than how they are conventionally used in everyday language or in classroom contexts. Any agent who generates observable data (e.g. by speaking, generating text or labels, performing an action to produce an effect) is a 'teacher' insofar as a learner can learn from the data. For instance, someone who presses a button to activate a toy is, in principle, a teacher if a learner observed this action and learned that pressing a button makes the toy go.

The key idea behind ISL is that learning can be especially powerful when both the teacher and the learner behave cooperatively (i.e. the teacher wants to help the learner learn, and the learner wants to learn from the teacher). Such cooperation requires commonsense psychology: a generative model of how others' internal states give rise to their observable actions. These internal states include mental states (e.g. beliefs, desires, goals) as well as the expected utilities (costs and rewards) of others' goal-directed actions [41]. The learner draws inverse inferences about the world states based on the observable behaviours of the teacher (e.g. demonstrations, verbal instruction, selecting examples or labels) by using commonsense psychology to consider the teachers' mental states and utilities; at the same time, the teacher selects or generates the best set of data (i.e. demonstrations, instructions, labels, examples) to help the learner learn by using their commonsense psychology to consider the learner's mental states and utilities. The process by which a cooperative teacher selects the data for the learner is referred to as pedagogical sampling [48,50,53], which can significantly reduce the amount of data (e.g. number of examples) the learner needs to converge on the correct hypothesis.

The consequences of such mutually cooperative interaction between the learner and the teacher have been demonstrated in humans, especially in young children, in the following ways:

(a) Humans as learners: inference and evaluation

In ISL, the learner's inferences from data provided by the teacher (e.g. demonstrations, examples of a concept) are modulated by what the learner knows about the teacher's mental states. This allows the learner to draw inferences that go beyond the observed data especially when the teacher is assumed to be knowledgeable and helpful [48,50]. Experiments with children [51,52,62,63] have shown that even infants and children make this inferential leap when demonstrations or examples of concepts indicate pedagogical sampling. For instance, when a teacher who claims to know all about a complex-looking toy and demonstrates just one interesting function of the toy (e.g. pressing a red button makes the toy light up), children not only learn that the red button turns on the light, but also 'go beyond the data' to infer that it is the *only* function of the toy; as a consequence, they explore the toy's other functions less than if the teacher were unfamiliar with the toy or interrupted partway through the demonstration. An understanding of the teacher's mental states is critical for this leap; if the toy had additional functions, a helpful, knowledgeable teacher would have demonstrated them.

Importantly, not all teachers are equally knowledgeable and helpful. A learner may encounter an ignorant or even a malicious (i.e. adversarial) teacher whose goal is to mislead or deceive. Learners, however, can shield themselves from such teachers by evaluating the data provided by the teacher against their own prior knowledge and update their mental model of the teacher; if the teacher is suspected to be ignorant or adversarial, the learner could adjust their future learning accordingly by placing less weight on data from the teacher or seek alternative sources of information. Human children are capable of such evaluation and selective learning even when the teacher technically provides accurate information; when a teacher demonstrates just one function of a toy while omitting other functions of the toy, children rate the teacher as less helpful compared to when the teacher demonstrates the toy's one and only function [64,65]. These evaluations can be modelled as probabilistic inferences about the quality of the teacher's

pedagogical sampling: the degree to which the teacher selected the best set of data for the learner given what the learner knows and wants [53].

(b) Humans as teachers: data-selection and communication

In ISL, a cooperative teacher considers the learner's mental states and expected utilities to select the data most helpful for the learner. Studies have shown that even young children can act as cooperative teachers who engage in pedagogical sampling; they go out of the way (i.e. incur extra costs of their own) to provide information that the learner does not know and wants to learn [66], but reasonably omit what the learner already knows or does not need [67]. For instance, when only 3 of 20 buttons on a toy play music (and the rest of the buttons are inert), children flexibly demonstrate either all 20 or just 3 buttons depending on what the learner already knows. When the learner doesn't know anything about the toy, they press all buttons (including the inert ones) because seeing just a few working buttons may lead the learner to infer that all buttons play music. However, if the learner already knows that only a few buttons play music (but doesn't know which ones do), children simply demonstrate the three working buttons [67]. Children also prioritize demonstrating a toy that would be harder for the learner to discover on their own (i.e. high discovery costs) over a toy that the learner can readily discover on their own [54]. Thus, even young children can act as a helpful teacher by considering a learner's knowledge, experiences and expected utilities.

Why do these findings matter for building socially intelligent machines? Currently, humans need a lot of expertise to provide input that is comprehensible to machines (e.g. prompt design). Likewise, machines need to be designed with a lot of care to produce output that is meaningful to humans (e.g. interpretable AI). What makes ISL useful as a case study is that the ways in which humans learn from and teach one another do not require formal training or expertise. Even young children can do it, both as learners and as teachers, because ISL is rooted in commonsense psychology.

3. Inferential social learning in machines

Given the importance of inferential social learning in humans, building machines capable of ISL would be a clearly desirable goal. Indeed, some recent work has emphasized ISL as a key property of autonomous agents that learn not only from the environment but also from human teachers [68]. Yet, it is also important to note that an account of human-to-human learning and communication cannot be blindly applied to any and all interactions between humans and machines. While communication between two human agents presupposes a shared conceptual space and inductive biases, machines do not come readily equipped with human concepts [29].

While implementing human-like ISL in machines still remains a major challenge, it is an exciting time to imagine a next generation of machines that are more socially intelligent than today's machines. Beyond simpler approaches like behavioural cloning [37], there is now a growing appreciation of how human input can improve machine learning [39,69,70] especially by inferring human values and goals to achieve better learning from human inputs and better alignment with human values (e.g. [40,71–73]).

While these trends signal exciting progress in AI, recent advances have also raised important questions. For instance, models trained with large-scale text data and human feedback (e.g. ChatGPT [9], GPT4 [74]) can learn from human inputs and even generate different outputs depending on the audience (e.g. 'explain photosynthesis to a 7-year-old child'). Are these machines capable of learning, teaching and communicating in ways that are premised on a shared causal understanding of other minds, as humans do in ISL? When are these abilities critical, and when are they not? In light of recent debates in AI about whether high-performing models we build are capable of human-like reasoning or just 'great memorization machines' (e.g. [10–12]),

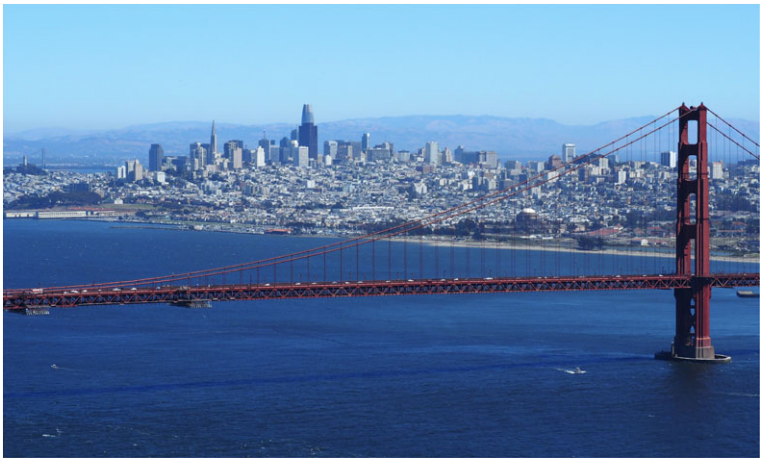

Figure 2. This image of a bridge near a city could be labelled or described differently depending on the context.

situating machines and humans in an ISL framework may help us clarify answers to these questions.

AI is a vast and fast-growing field. Thus our goal is not to suggest specific solutions for a particular problem, but to situate humans and machines in the ISL framework to identify new questions and suggest promising directions that emerge from this picture. In what follows, we envision machines that learn from humans (i.e. machines as learners, §3(a)) and help humans learn (i.e. machines as teachers, §3(b)).

(a) Towards socially intelligent machines that learn from humans

In §2, we provided key characteristics of an inferential social learner in humans (flexible inference from data, sensitivity to the quality of data). In a nutshell, a learner capable of ISL capitalizes on the fact that humans, as teachers, speak and act differently depending on their mental states and the intended audience; that is, the teacher's behaviours come from their mental states, modulated by their communicative intent, and the learner benefits from considering these factors. We can draw parallels to these characteristics by considering a machine learning model as the learner and the source of data (i.e. humans who generated the training data) as the teacher.

(i) Understanding the knowledge and intentions of human teachers (data-generators)

Suppose you asked others to describe the image in figure 2. While some might refer to the 'Golden Gate Bridge' (i.e. the prominent structure featured in the foreground) or 'San Francisco' (i.e. the city in the background), those who do not recognize the bridge or the cityscape may use other terms to describe the salient and distinct features of this image (e.g. bridge, city). Furthermore, even the same person could describe the image very differently depending on the constraints on the output (e.g. a single label or free response?), the audience (e.g. a child or an adult?) and the purpose (e.g. a description for a potential tourists?). Thus, data generated by human 'teachers'—whether it be labels, demonstrations, or large amounts of text—reflect their goals, knowledge, intentions, and what they know about the potential recipient. If a machine learner can consider the human teacher's mental states and the generative process that gave rise to the choice of labels or descriptions, the machine learner can potentially 'go beyond the data' to increase sampling efficiency, or even protect itself from potentially harmful sources of information by evaluating the quality of data.

From this perspective, developing ISL-inspired machines will require richer datasets that reflect a broader range of human intentions and knowledge. While some of the most influential datasets in computer vision have often included human-generated labels, annotations, or even questions [75–77], humans in these datasets were given minimal prompts with little context about how the labels would be used. Thus, these datasets contain human-generated outputs that have high convergence across individuals (strength) but lack the variance in communicative intent that are typical of human interactions (weakness). A promising way forward may be to build datasets that expose the importance of communicative intentions and social context. Human demonstrations are already a valuable source of data for training machines to perform long-range planning tasks in complex physical environments (e.g. household tasks such as cleaning a bedroom [78]). While these demonstrations may not necessarily be communicative in nature, they can still vary depending on the expertise and knowledge of the demonstrator (professional cleaner or a novice?) or the intended observer (humans or robots?). More generally, datasets displaying greater variation in human intentions can facilitate the development of more powerful machines that leverage human goals and preferences to learn more effectively from humans and even coordinate their own actions with humans [72,79].

(ii) Beyond rich datasets: ISL in foundation models

A very different example of a rich dataset is the internet itself. Large language models (LLMs), or foundation models more generally [80], are pre-trained deep neural network models that are increasingly integrated with other model components for various tasks, ranging from answering simple questions to writing codes or composing an essay (e.g. GPT3 [81], GPT4 [74] and PALM [82]). These models are trained on internet-scale text data generated by humans, originally intended to be consumed by other humans, often with a clear communicative intent or even with a specific audience in mind. While these models are trained to learn patterns in text without any explicit consideration of such communicative intent in the training process, recent work suggests that using chain-of-thought prompts that reflect the latent human communicative intent can improve their performance [83,84]. Furthermore, recent work also raises the possibility that language models, in principle, could learn representations that can correspond to human mental states, such as beliefs, desires, and goals [85]. Recent debates about whether LLMs have genuine, human-like linguistic competence [10,11] or Theory of Mind [12,86] are also relevant to the possibility of large models acquiring a model of the human mind.

Particularly notable examples involve reinforcement learning from human feedback (RLHF) where models are further fine-tuned using explicit human feedback, as in the case of InstructGPT and ChatGPT [9,70]. Here, one can imagine the human rater's role as both a learner and a teacher; the human rater is a learner who evaluates the quality and helpfulness of the model's output, but to the extent that their feedback is used for training, the human is also a teacher. Current approaches use human ratings (e.g. rank-order of prompts) as part of the objective function, reflecting a broader trend in the cognitive neuroscience literature that treats social feedback as a simple reward-predictive cue [87]. However, when humans provide evaluative feedback to other humans, they often do so with a communicative intent (e.g. help a student improve), with different expectations and standards depending on who is being evaluated on what kind of task. This means that even the same response to a prompt may be evaluated differently depending on the context. By placing machines and humans in ISL framework, we can envision learning from human feedback that capitalizes on a broad range of human raters' mental states and values. For instance, a machine learner capable of ISL could learn more from less data based on a few 'best' and 'worst' examples chosen by a human rater (teacher), and prioritize learning from certain raters depending on their reliability and evaluative standards.

As noted above, a socially intelligent machine that is capable of considering the intentions of the data source (teacher) should also be capable of learning to detect (and avoid further learning from) a malicious source of data. In an online learning setting where a production model might be fine-tuned as it receives new data (e.g. sudden surge of a keyword in search), it is crucial

to decide whether to adopt to a distributional change (e.g. breaking news) or ignore as outliers (e.g. adversarial attacks). Incidents where new AI (e.g. Tay chatbot, [88]) is released in the wild and taken down immediately due to problems that arose from learning from online data are examples of failures to consider the human intentions that gave rise to the observed data. More generally, being able to consider the role of humans and the quality of their inputs in the training process would be particularly critical as various foundation models—trained by large-scale data from difference sources with varying degree of quality and fidelity over the course of a long time—become more influential in our society.

Looking further ahead, developing machines that are capable of understanding the latent variables (e.g. mental states) that give rise to human behaviours is deeply relevant to the alignment problem, which refers to the challenge of steering AI systems towards human goals and interests [89]. For instance, if a robot starts cleaning the house while a human is trying to take a nap, a socially intelligent robot would not take the human's negative feedback (e.g. 'don't do that again!') as a blanket ban on vacuuming; it would understand that the human wants to nap, that the vacuum noise is preventing him from getting it, and it should stop and resume the task when the human is awake or away.[2] In fact, even understanding what 'that' in the human command refers to is a non-trivial problem. While this is a toy example, the general ability to understand the *why's* in human users' inputs and responding appropriately will not only require advances in AL/ML but also in cognitive science, as we still do not fully understand how humans are capable of understanding such commands.

(b) Towards socially intelligent machines that help humans learn

The prevailing current approach in AI is to think about machines as learners that must be trained with large human-generated datasets to be useful. However, as machines become more powerful learners, they can also play increasingly important roles as teachers that help humans learn. Here, we consider two ways in which machines can serve as teachers in a broad sense. First, machines could better translate its predictions, decisions and discoveries in ways that are understandable to humans; second, machines could teach human learners in educational settings.

(i) Machines that can expand human knowledge

Today's AI systems can already perform tasks better than human experts in some domains [13,14]. However, the rationale behind the predictions made by the most performant of these systems remains mostly incomprehensible to humans, motivating the need to develop strategies to probe how these systems arrive at their decisions, and how to translate them in a way that is understandable to humans [29,91]. Progress in developing more interpretable AI systems may not only promote safer use of these systems, but also open up the possibility for them to genuinely expand human knowledge.

For example, consider AlphaGo, an AI system trained to play the strategy board game Go, a game noted for its substantial complexity. In a high-profile match-up between AlphaGo [13] and human world champion Lee Sedol in 2016, AlphaGo generated a surprising move ('move 37' [92]), which turned out to be critical to its historic victory over Lee. This move confused even expert human players, fuelling excitement about the possibility of expanding the suite of effective Go strategies known to humans. However, there are many challenges in making AlphaGo's knowledge accessible to humans. First, while the moves AlphaGo makes are directly observable, the internal processes giving rise to these moves are not. Second, given the many differences between AlphaGo and humans, it is likely that the latent representations that AlphaGo has acquired to play Go are different from the latent representations that human players use when playing the same game.

Current approaches to bridging these observability and representational gaps include building 'inherently interpretable' structure into a model during training [93] or by introducing such

[2]This example was mentioned in The Book of Why: The New Science of Cause and Effect by Pearl & Mackenzie [90].

structure after training (i.e. *post hoc* methods [15,–17,94]). Some common ways of introducing 'structure' into inherently interpretable models are to define rules [95], provide prototypical examples [19,96], inducing causal structure [97] and more recently, employ human-interpretable concepts [17,98,101] rather than 'raw' model feature representations [15,16,94]. The intuition behind using *concepts* is to bring model representations into closer alignment with the vocabulary that humans use. For instance, a machine could use part concepts like 'feather' and 'wings' to describe why it recognized a 'bird' in an image rather than referring to individual pixels in the image. This approach has gained traction especially in highly specialized domains (e.g. medicine [102–105]), as it allows a mapping between the components of an AI system's decision processes and the concept labels that humans understand (e.g. translating a model's predictions in ways that make sense to medical professionals in clinical settings, see [102]).

These approaches, however, use human-provided concept labels for each specific domain, rather than taking a general approach to inferring human-like concept-based representations that can be applied in any arbitrary domain. As such, they rely on assumptions about human concepts that may not always be justified. For example, similarity measures in inherently interpretable models may differ substantially from how humans reason about similarity [106,107]). To the extent that the latent representations used by models to reason about these concepts differ from those actually used by humans, relying on model-based similarities alone may introduce various distortions and biases that limit human interpretability. More generally, a key issue in interpretability lies in understanding *why* humans describe their observations and reasoning processes using certain concepts and not others, or what information humans might find most useful and relevant to their goals.

Beyond bringing machines closer to humans by increasing interpretability, another avenue forward may be to embrace the possibility that humans will need to acquire fundamentally new concepts in order to understand what highly capable AI systems have learned. After all, examples of radical forms of conceptual change abound in how children learn [108,109]. For instance, children take years to acquire numerical concepts like 'one,' 'two' and 'three' [110]; before a child acquires an adult-like concept of numbers, how the child understands information about quantity may be opaque to an adult, and vice versa. Expanding what humans know by learning from machines may entail a similar sort of radical conceptual change. An intriguing possibility is that such changes might be driven by the same mechanisms that underlie conceptual change and cognitive development in humans.

(ii) Machines that support formal education

A major challenge confronting any teacher in a classroom is selecting the right kind of information to facilitate student learning; this is a real-world version of the problem that computational models of pedagogical reasoning try to address [48]. Crucially, as we have seen in the above descriptions of ISL, the value of information for a given student depends not only on the predetermined content of a particular lesson but also on the students' knowledge, skills, and goals (i.e. the student's mental states and utilities). Understanding these internal qualities of a student requires non-trivial forms of inference that often must be made from a limited amount of data (e.g. a few submissions by the student on a coding assignment). Moreover, even when a teacher draws accurate inferences about an individual student's knowledge state, it is prohibitively costly in time and resources for a single teacher to design personalized lesson plans for all of their students.

As such, we see major opportunities to develop AI systems that support teachers by both helping them keep track of student learning, as well as design and administer learning activities that are tailored to individual student needs. For instance, while some students may benefit more from concrete examples that illustrate a certain concept, other students may benefit from by being provided with the abstract principle [111]. In these scenarios, building AI systems that are able to leverage student output (e.g. responses to open-ended questions) to accurately model how a specific student's beliefs would change given an example or principle is more likely to offer more useful pedagogical recommendations [112]. In addition, because what a student knows

is expected to change across days, weeks, and months, it is important to develop approaches that integrate student behaviour over these multiple timescales when determining what kind of feedback or learning activity is likely to be helpful to a student. As such, developing AI systems that help human teachers keep track of individual student learning trajectories may be especially impactful for educators.

There are already efforts underway to develop AI systems that meet some of these challenges. For example, there is a growing body of work leveraging massive datasets from many students to infer which concepts an individual student might be struggling with [113–116]. In addition, there are continually improving systems that aim to make good recommendations about the sequence of lessons that are appropriate for individual students [38,117], as well as ones that aim to provide real-time feedback to students as they engage with new material [118,119]. However, in more naturalistic pedagogical settings where social feedback produced by current machines has been compared directly with that provided by human teachers, it appears that current machine teachers still underperform on several key measures of pedagogical success [120]. Nevertheless, we are optimistic about the prospects for applications of AI in formal education, especially AI systems that embrace the importance of building genuine mental models of human learners.

4. Humans as learners and teachers: studying how humans think about machines

Social learning and teaching in humans involve much more than a teacher feeding data to a learner; it is built upon a mutual understanding between a teacher and a learner. Therefore, efforts to develop machines that learn from, teach, and communicate with humans cannot succeed without understanding how *humans* think about machine minds. Research that investigates the mental models that humans use to understand machine behaviour is barely in its infancy, but there are already signs of exciting progress.

(a) Intuitive theory of AI: how humans build mental models of machine minds

The human mind is not inherently equipped to represent or reason about AI. While humans have distinct cognitive and neural mechanisms for perceiving and representing 'living things' and their minds [121–124] versus inanimate objects and physical events [125–127], when objects move like animate beings, humans (even infants) tend to attribute agency and mental capacities to these entities [128–130]. From this perspective, AI presumably occupies an intermediate position between objects and agents; some AI systems resemble human appearance and are capable of self-locomotion (e.g. humanoid robots) while others clearly look like objects (e.g. smart speakers) but are still capable of processing language. How do humans make sense of these entities?

Recent research suggests that the kinds of mental capacities humans (especially young children) attribute to AI depends not only on their human-likeness in appearance, but more importantly on how they behave and respond to human input [131–133], reflecting that AI indeed occupies this in-between space. One possibility is that humans construct an *intuitive theory of AI*, a system of beliefs about what it is and how it behaves, just as how humans' beliefs about objects and agents have been characterized as intuitive theories [134,135]. Rather than being built from scratch, however, its initial components may come from their knowledge about objects and agents, and become gradually revised and refined based on their experience with AI systems. Thus, contemporary theories of how humans build such theories [43,44,134] as well as computational approaches to studying social cognition [41,46,87]) could be applied fruitfully to the question of how humans construct an intuitive theory of machine minds.

More generally, investigating how humans represent a diverse array of machines and machine minds across ages, cultures and degree of experience with AI is an exciting avenue for research, and would provide important grounds for thinking about human–machine communication. Collaborations between cognitive science and ML/AI can also inform specific questions about

the degree to which humans regard machines as a socially intelligent *agent* as opposed to simply a useful device (i.e. an *object*). What are the kinds of errors that are diagnostic of machine outputs, and how are they different from the ones only humans would make? How rich of a mental model do people spontaneously ascribe to systems that make errors that seem qualitatively different from their own? Answering such questions rigorously is challenging due to their multidisciplinary nature, but crucial for both fields; these efforts can help develop safer and more trustworthy AI, while also having profound implications for how humans interact with machines and evaluate its errors. We further expand on this point below.

(b) Moral evaluation of machines

The question of how humans think about machine minds is particularly urgent in high-stakes settings where humans and machines must quickly coordinate their behaviours. For example, autonomous vehicles must be able to accurately infer the goals and plans of other agents operating other vehicles on the road in order to safely navigate such environments [71,79]. However, when accidents involving autonomous vehicles invariably occur [136,137], these situations are precisely where humans' intuitive theory of AI will have real-world relevance: to the extent that humans expect machines as an agent with a goal to safely operate a passenger vehicle (rather than a device that simply follows rules), it may also be expected to bear moral responsibility for the split-second decisions it makes [138–140].

Although both human and machine minds are vulnerable to making errors, current AI systems make systematically different errors than humans do in several domains that pertain to safe driving, including visual object categorization [6,141] and physical scene understanding [142]. Nevertheless, these behavioural differences do not necessarily imply that machine latent representations are entirely opaque to human observers. Recent work in cognitive science has found that humans are capable of quickly building reasonable context-specific expectations about how these machines will behave in visual recognition [4,143] and physical prediction tasks [144]. These demonstrations suggest that ordinary people build and use mental models of machines, even when (or *especially* when) these machines make surprising errors. Given that people tend to blame human drivers more than their automated cars when either of them make a mistake [139], studying how people conceptualize various forms of AI agents and the impact of different conceptualizations on moral judgement continue to be an important line of inquiry.

On another note, even though we anticipate meaningful progress towards human–machine mutual understanding in the coming years, it will be important to continue to manage the expectations that people have about these systems. After all, machines are not humans, and even once they approach human-level abilities in many domains, even in their social inference and communicative abilities, there will still likely be contexts where there remain important gaps between humans and any given AI system. It may thus be critical for socially intelligent AI systems to be sensitive to human users' expectations and convey their uncertainty in a manner that humans will appreciate.

5. Looking ahead, working together: AI and cognitive science

So far, we have argued that we need socially intelligent machines that can understand and be understood by humans. We envision that these next-generation machines learn in smarter ways by considering the role of humans (both in training with massive datasets and learning from small amounts of human input), and help humans learn not only by serving as effective teachers but also by making their decisions more understandable by human users. This fundamentally cooperative nature of ISL may also help minimize the potential risks of machines with powerful social intelligence. To achieve this, we need to go beyond the engineering-centric approach, towards a science-centric approach [145–147], and study machines as targets of scientific study. These advances will not be possible without both fields working together. While machine learning/AI research needs insights from cognitive science to build machines that can think about

humans, these efforts must be complemented by scientific studies of how humans think about machines.

An interesting open challenge that goes beyond the scope of this paper is *measuring* our progress towards more socially intelligent AI. A reasonable starting point could be to leverage techniques already used in cognitive science and psychology to study human intelligence. For instance, we may need to develop new tools—e.g. tests of AI behaviour with high measurement reliability and construct validity—to thoroughly probe how these systems work. In particular, experimental methods developed for studying the human mind [148] may prove to be useful for studying the 'minds' of machines. Just as such tools are used in cognitive science and psychology to study how the human mind works, we anticipate the 'cognitive science of AI' emerging as an important research area that aims to understand the degree to which an AI system has acquired an understanding of various core concepts [142,145].

Taking such an approach has been particularly useful for advancing our understanding of intelligence in children versus adults (e.g. cognitive development) as well as humans versus other species (e.g. comparative cognition), where the mechanistic causes of behavioural changes (or differences) are not directly observable. Taking a similar approach to make well-controlled comparisons between different kinds of AI systems might offer generalizable insights about the *mechanistic* causes of their behaviours. Some recent benchmarks inspired by infant social cognition [149,150] represent the first steps in this direction, attempting to measure machines' ability to predict an agent's behaviour with respect to its goals, preferences, and action efficiency in minimal settings. As the field moves towards machines that are actually able to interact with humans (i.e. embodied robots), it will become increasingly critical to have benchmarks that test complex machine behaviours in more realistic settings [78,151]. Ultimately, coordinating efforts across all of these fields will be critical for developing a unified set of toolkits for studying both natural (i.e. human) and artificial intelligence.

Going beyond characterizations of intelligence that focuses on an isolated agent, we anticipate that social intelligence will take a central place in these endeavours. Our contemporary understanding of the human mind has been largely shaped by the 'cognitive revolution' in the 1950s. This was a response to the behaviourist ideas that treated the mind as a black box and focused only on what goes in (input/data) and what comes out (output/behaviour), characterizing learning as conditioned responses or strengthening of associations. However, despite substantial advances that have thrown more light into this black box, it has taken decades for theories of human learning to embrace the centrality of social cognition in how humans learn and think. Our hope is that AI research can avoid retracing the long detour by treating social intelligence as an integral part of what it means to develop smarter machines.

Data accessibility. This article has no additional data.

Authors' contributions. H.G.: conceptualization, writing—original draft, writing—review and editing; J.F.: conceptualization, writing—original draft, writing—review and editing; B.K.: conceptualization, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This article was supported by: NSF CNS-2120095, NSF BCS-2019567, and McDonnell Scholars Award (HG) and NSF CAREER-2047191 and ONR Science of Autonomy (JF).

Acknowledgements. We are grateful for two anonymous reviewers as well as Robert Geirhos and Natalia Vélez for helpful comments and feedback.

References

1. Tomasello M, Kruger AC, Ratner HH. 1993 Cultural learning. *Behav. Brain Sci.* **16**, 495–552. (https://core.ac.uk/download/pdf/85210124.pdf)
2. Henrich J. 2015 *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton, NJ: Princeton University Press.

3. Maslej N *et al.* 2023 *The AI index 2023 annual report. AI index steering committee*. Stanford, CA: Stanford Institute for Human-Centered AI.
4. Zhou Z, Firestone C. 2019 Humans can decipher adversarial images. *Nat. Commun.* **10**, 1–9. (doi:10.1038/s41467-019-08931-6)
5. Lepori MA, Firestone C. 2022 Can you hear me *now*? Sensitive comparisons of human and machine perception. *Cogn. Sci.* **46**, e13191. (doi:10.1111/cogs.13191)
6. Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ. 2018 Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* **38**, 7255–7269. (doi:10.1523/JNEUROSCI.0388-18.2018)
7. Schrimpf M *et al.* 2020 Brain-score: which artificial neural network for object recognition is most brain-like? *BioRxiv*, p. 407007.
8. Caucheteux C, King JR. 2022 Brains and algorithms partially converge in natural language processing. *Commun. Biol.* **5**, 1–10. (doi:10.1038/s42003-022-03036-1)
9. ChatGPT. See https://openai.com/blog/chatgpt/ (accessed 13 December 2022).
10. Sap M, LeBras R, Fried D, Choi Y. 2022 Neural theory-of-mind? on the limits of social intelligence in large LMs. (https://arxiv.org/abs/2210.13312)
11. Mahowald K, Ivanova AA, Blank IA, Kanwisher N, Tenenbaum JB, Fedorenko E. 2023 Dissociating language, thought in large language models: a cognitive perspective. (https://arxiv.org/abs/2301.06627)
12. Ullman T. 2023 Large language models fail on trivial alterations to theory-of-mind tasks. (https://arxiv.org/abs/2302.08399)
13. Silver D *et al.* 2016 Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489. (doi:10.1038/nature16961)
14. Jumper J *et al.* 2021 Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589. (doi:10.1038/s41586-021-03819-2)
15. Lundberg SM, Lee SI. 2017 A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, vol. 31. New York, NY: Curran Associates, Inc.
16. Sundararajan M, Taly A, Yan Q. 2017 Axiomatic attribution for deep networks. In *Proc. of the 34th Int. Conf. on Machine Learning*. New York, NY: Curran Associates, Inc.
17. Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F. 2018 Interpretability beyond feature attribution: quantitative testing with concept activation vectors (tcav). In *Int. Conf. on Machine Learning*, pp. 2668–2677. New York, NY: PMLR.
18. Angelino E, Larus-Stone N, Alabi D, Seltzer M, Rudin C. 2017 Learning certifiably optimal rule lists for categorical data. *J. Mach. Learn. Res.* **18**, 8753–8830. (doi:10.1145/3097983.3098047)
19. Chen C, Li O, Tao D, Barnett A, Rudin C, Su JK. 2019 This looks like that: deep learning for interpretable image recognition. In *Advances in neural information processing systems* (eds H Wallach, H Larochelle, A Beygelzimer, F d'Alché Buc, E Fox, R Garnett), vol. 32. New York, NY: Curran Associates, Inc.
20. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. 2017 Grad-cam: visual explanations from deep networks via gradient-based localization. In *Proc. of the IEEE Int. Conf. on Computer Vision*, pp. 618–626. Washington, DC: IEEE.
21. Vasconcelos H, Jörke M, Grunde-McLaughlin M, Krishna R, Gerstenberg T, Bernstein MS. 2022 When do XAI methods work? A cost-benefit approach to human-AI collaboration. In *CHI Workshop on Trust and Reliance in AI-Human Teams*. New Orleans: ACM. See https://arxiv.org/pdf/2212.06823.pdf.
22. Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Wortman Vaughan J. 2020 Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proc. of the 2020 CHI Conf. on Human Factors in Computing Systems*, pp. 1–14. New York NY: ACM.
23. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Vaughan JW, Wallach H. 2021 Manipulating and measuring model interpretability. In *Proc. of the 2021 CHI Conf. on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021*, pp. 1–52. New York, NY: ACM.
24. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. 2018 Sanity checks for saliency maps. *Adv. Neural Inform. Process. Syst.* **31**, 9525–9536.

25. Adebayo J, Muelly M, Abelson H, Kim B. 2022 Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *Proc. of the 10th Int. Conf. on Learning Representations*, Online, 25–29 April 2022. Appleton, WI: ICLR.
26. Alvarez-Melis D, Jaakkola TS. 2018 On the robustness of interpretability methods. In *Proc. of the ICML 2018 Workshop on Human Interpretability in Machine Learning*. New York, NY: Curran Associates Inc.
27. Adebayo J, Muelly M, Liccardi I, Kim B. 2020 Debugging tests for model explanations. In *34th Conf. on Neural Information Processing Systems, Vancouver, Canada, 6–12 December 2020*. New York, NY: Curran Associates, Inc.
28. Bilodeau B, Jaques N, Koh PW, Kim B. 2022 Impossibility theorems for feature attribution. (https://arxiv.org/abs/2212.11870)
29. Kim B. 2022 Beyond interpretability: developing a language to shape our relationships with AI. In *10th Int. Conf. on Learning Representations*, Online, 25–29 April 2022. Appleton WI: ICLR.
30. Gweon H. 2021 Inferential social learning: cognitive foundations of human social learning and teaching. *Trends Cogn. Sci.* **25**, 896–910. (doi:10.1016/j.tics.2021.07.008)
31. Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. 2016 Building machines that learn and think like people. *Behav. Brain Sci.* **40**, 1–101. (doi:10.1017/S0140525X16001837)
32. Breazeal C, Scassellati B. 2002 Robots that imitate humans. *Trends Cogn. Sci.* **6**, 481–487. (doi:10.1016/S1364-6613(02)02016-8)
33. Breazeal C, Dautenhahn K, Kanda T. 2016 *Social robotics*. Springer Handbook of Robotics, pp. 1935–1972. Berlin, Germany: Springer.
34. Perolat J *et al.* 2022 Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science* **378**, 990–996. (doi:10.1126/science.add4679)
35. Bakhtin A *et al.* MFARDT (FAIR)†. 2022 Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science* **378**, 1067–1074. (doi:10.1126/science.ade9097)
36. Abbeel P, Ng AY. 2004 Apprenticeship learning via inverse reinforcement learning. In *Proc. of the Twenty-First Int. Conf. on Machine learning*, p. 1.
37. Torabi F, Warnell G, Stone P. 2018 Behavioral cloning from observation. (https://arxiv.org/abs/1805.01954)
38. Zhu X. 2015 Machine teaching: an inverse problem to machine learning and an approach toward optimal education. In *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 29. Palo Alto, CA: AAAI Press.
39. Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D. 2017 Deep reinforcement learning from human preferences. *Adv. Neural Inform. Process. Syst.* **30**, 4302–4310.
40. Hadfield-Menell D, Russell SJ, Abbeel P, Dragan A. 2016 Cooperative inverse reinforcement learning. *Adv. Neural Inform. Process. Syst.* **29**, 3916–3924.
41. Jara-Ettinger J, Gweon H, Schulz LE, Tenenbaum JB. 2016 The Naïve Utility Calculus: computational Principles Underlying Commonsense Psychology. *Trends Cogn. Sci.* **20**, 589–604. (doi:10.1016/j.tics.2016.05.011)
42. Oaksford M, Chater N. 2007 *Bayesian rationality: the probabilistic approach to human reasoning*. Oxford, UK: Oxford University Press.
43. Tenenbaum JB, Kemp C, Goodman ND. 2011 How to grow a mind: statistics, structure, and abstraction. *Science* **331**, 1279–1285. (doi:10.1126/science.1192788)
44. Gopnik A. 2012 Scientific thinking in young children: theoretical advances, empirical research, and policy implications. *Science* **337**, 1623–1627. (doi:10.1126/science.1223416)
45. Schulz L. 2012 The origins of inquiry: inductive inference and exploration in early childhood. *Trends Cogn. Sci.* **16**, 382–389. (doi:10.1016/j.tics.2012.06.004)
46. Baker CL, Jara-Ettinger J, Saxe R, Tenenbaum JB. 2017 Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Publish. Group* **1**, 1–10. (doi:10.1038/s41562-017-0064)
47. Jara-Ettinger J. 2019 Theory of mind as inverse reinforcement learning. *Curr. Opin. Behav. Sci.* **29**, 105–110. (doi:10.1016/j.cobeha.2019.04.010)
48. Shafto P, Goodman ND, Griffiths TL. 2014 A rational account of pedagogical reasoning: teaching by, and learning from, examples. *Cognit. Psychol.* **71**, 55–89. (doi:10.1016/j.cogpsych.2013.12.004)

49. Goodman ND, Frank MC. 2016 Pragmatic language interpretation as probabilistic inference. *Trends Cogn. Sci.* **20**, 818–829. (doi:10.1016/j.tics.2016.08.005)
50. Wang P, Wang J, Paranamana P, Shafto P. 2020 A mathematical theory of cooperative communication. *Adv. Neural Inf. Process. Syst.* **33**, 17 582–17 593.
51. Gweon H, Tenenbaum JB, Schulz LE. 2010 Infants consider both the sample and the sampling process in inductive generalization. *Proc. Natl Acad. Sci. USA* **107**, 9066–9071. (doi:10.1073/pnas.1003095107)
52. Bonawitz E, Shafto P, Gweon H, Goodman ND, Spelke E, Schulz L. 2011 The double-edged sword of pedagogy: instruction limits spontaneous exploration and discovery. *Cognition* **120**, 322–330. (doi:10.1016/j.cognition.2010.10.001)
53. Bass I, Bonawitz E, Hawthorne-Madell D, Vong WK, Goodman ND, Gweon H. 2022 The effects of information utility and teachers' knowledge on evaluations of under-informative pedagogy across development. *Cognition* **222**, 104999. (doi:10.1016/j.cognition.2021.104999)
54. Bridgers S, Jara-Ettinger J, Gweon H. 2020 Young children consider the expected utility of others' learning to decide what to teach. *Nat. Hum. Behav.* **4**, 144–152. (doi:10.1038/s41562-019-0748-6)
55. Lyons DE, Young AG, Keil FC. 2007 The hidden structure of overimitation. *Proc. Natl Acad. Sci. USA* **104**, 19 751–19 756. (doi:10.1073/pnas.0704452104)
56. Legare CH, Nielsen M. 2015 Imitation and innovation: the dual engines of cultural learning. *Trends Cogn. Sci.* **19**, 688–699. (doi:10.1016/j.tics.2015.08.005)
57. Harris PL. 2012 *Trusting what you're told: how children learn from others*. Cambridge, MA: Harvard University Press.
58. Gergely G, Bekkering H, Kiraly I. 2002 Rational imitation in preverbal infants. *Nature* **415**, 755. (doi:10.1038/415755a)
59. Koenig MA, Harris PL. 2005 The role of social cognition in early trust. *Trends Cogn. Sci.* **9**, 457–459. (doi:10.1016/j.tics.2005.08.006)
60. Sobel DM, Kushnir T. 2013 Knowledge matters: how children evaluate the reliability of testimony as a process of rational inference. *Psychol. Rev.* **120**, 779–797. (doi:10.1037/a0034191)
61. Griffiths TL. 2020 Understanding human intelligence through human limitations. *Trends Cogn. Sci.* **24**, 873–883. (doi:10.1016/j.tics.2020.09.001)
62. Xu F, Tenenbaum J. 2007 Word learning as Bayesian inference. *Psychol. Rev.* **114**, 245–272. (doi:10.1037/0033-295X.114.2.245)
63. Shneidman L, Gweon H, Schulz LE, Woodward AL. 2016 Learning from others and spontaneous exploration: a cross-cultural investigation. *Child Dev.* **87**, 723–735. (doi:10.1111/cdev.12502)
64. Gweon H, Pelton H, Konopka JA, Schulz LE. 2014 Sins of omission: children selectively explore when teachers are under-informative. *Cognition* **132**, 335–341. (doi:10.1016/j.cognition.2014.04.013)
65. Gweon H, Asaba M. 2018 Order matters: children's evaluation of underinformative teachers depends on context. *Child Dev.* **89**, e278–e292. (doi:10.1111/cdev.12825)
66. Gweon H, Schulz L. 2019 From exploration to instruction: children learn from exploration and tailor their demonstrations to observers' goals and competence. *Child Dev.* **90**, e148–e164. (doi:10.1111/cdev.13059)
67. Gweon H, Shafto P, Schulz L. 2018 Development of children's sensitivity to overinformativeness in learning and teaching. *Dev. Psychol.* **54**, 2113–2125. (doi:10.1037/dev0000580)
68. Sigaud O, Akakzia A, Caselles-Dupré H, Colas C, Oudeyer PY, Chetouani M. 2022 Towards teachable autotelic agents. *IEEE Trans. on Cognitive and Developmental Systems*. Washington, DC: IEEE.
69. Krishna R, Lee D, Fei-Fei L, Bernstein MS. 2022 Socially situated artificial intelligence enables learning from human interaction. *Proc. Natl Acad. Sci. USA* **119**, e2115730119. (doi:10.1073/pnas.2115730119)
70. Ouyang L *et al.* 2022 Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35, New Orleans, LA, 28 November - 9 December 2022*. New York, NY: Curran Associates, Inc.

71. Sadigh D, Sastry S, Seshia SA, Dragan AD. 2016 Planning for autonomous cars that leverage effects on human actions. In *Robotics science and systems*, vol. 2, pp. 1–9. Ann Arbor, MI: Robotics Science and Systems.
72. Palan M, Landolfi NC, Shevchuk G, Sadigh D. 2019 Learning reward functions by integrating human demonstrations and preferences. (https://arxiv.org/abs/1906.08928)
73. Fisac JF *et al.* 2020 Pragmatic-pedagogic value alignment. In *Robotics Research: The 18th Int. Symp. ISRR*, pp. 49–57. Berlin, Germany: Springer.
74. OpenAI. 2023 GPT-4 technical report.
75. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. 2009 Imagenet: a large-scale hierarchical image database. In *2009 IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 248–255. Washington, DC: IEEE.
76. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. 2014 Microsoft coco: common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proc., Part V 13*, pp. 740–755. Berlin, Germany: Springer.
77. Rajpurkar P, Zhang J, Lopyrev K, Liang P. 2016 Squad: 100 000+ questions for machine comprehension of text. (https://arxiv.org/abs/1606.05250)
78. Srivastava S *et al.* 2022 Behavior: benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conf. on Robot Learning*, pp. 477–490. Brookline, MA: PMLR.
79. Toghi B, Valiente R, Sadigh D, Pedarsani R, Fallah YP. 2022 Social coordination and altruism in autonomous driving. *IEEE Trans. Intell. Transp. Syst.* **23**, 24 791–24 804. (doi:10.1109/TITS.2022.3207872)
80. Bommasani R *et al.* 2021 On the opportunities and risks of foundation models. (https://arxiv.org/abs/2108.07258)
81. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. 2019 Language models are unsupervised multitask learners.
82. Chowdhery A *et al.* 2022 Scaling language modeling with pathways. (https://arxiv.org/abs/2204.02311)
83. Wei J, Wang X, Schuurmans D, Bosma M, Chi E, Le Q, Zhou D. 2022 Chain of thought prompting elicits reasoning in large language models. (https://arxiv.org/abs/2201.11903)
84. Prystawski B, Thibodeau P, Goodman N. 2022 Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. (https://arxiv.org/abs/2209.08141)
85. Andreas J. 2022 Language models as agent models. (https://arxiv.org/abs/2212.01681)
86. Kosinski M. 2023 Theory of mind may have spontaneously emerged in large language models. (https://arxiv.org/abs/2302.02083)
87. Vélez N, Gweon H. 2021 Learning from other minds: an optimistic critique of reinforcement learning models of social learning. *Curr. Opin. Behav. Sci.* **38**, 110–115. (doi:10.1016/j.cobeha.2021.01.006)
88. Tay (bot) from Microsoft. See https://en.wikipedia.org/wiki/Tay_(bot) (accessed 16 December 2022).
89. Alignment problem. See https://en.wikipedia.org/wiki/AI_alignment (accessed 16 December 2022).
90. Pearl J, Mackenzie D. 2018 *The book of why: the new science of cause and effect*. New York, NY: Basic Books.
91. McGrath T, Kapishnikov A, Tomašev N, Pearce A, Wattenberg M, Hassabis D, Kim B, Paquet U, Kramnik V. 2022 Acquisition of chess knowledge in alphazero. *Proc. Natl Acad. Sci. USA* **119**, e2206625119. (doi:10.1073/pnas.2206625119)
92. In two moves, AlphaGo and Lee Sedol redefined the future. See www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/ (accessed 13 December 2022).
93. Rudin C. 2019 Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215. (doi:10.1038/s42256-019-0048-x)
94. Frosst N, Hinton G. 2017 Distilling a neural network into a soft decision tree.
95. Wu X *et al.* 2008 Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **14**, 1–37. (doi:10.1007/s10115-007-0114-2)

96. Kim B, Rudin C, Shah JA. 2014 The Bayesian case model: a generative approach for case-based reasoning and prototype classification. *Adv. Neural Inform. Process. Syst.* **27**, 1952–1960.
97. Geiger A, Wu Z, Lu H, Rozner J, Kreiss E, Icard T, Goodman N, Potts C. 2022 Inducing causal structure for interpretable neural networks. In *Int. Conf. on Machine Learning*, pp. 7324–7338. New York, NY: PMLR.
98. Ghandeharioun A, Kim B, Li CL, Jou B, Eoff B, Picard RW. 2021 Dissect: disentangled simultaneous explanations via concept traversals. (https://arxiv.org/abs/2105.15164)
99. Yeh CK, Kim B, Arik S, Li CL, Pfister T, Ravikumar P. 2020 On completeness-aware concept-based explanations in deep neural networks. *Adv. Neural Inf. Process. Syst.* **33**, 20 554–20 565.
100. Chen Z, Bei Y, Rudin C. 2020 Concept whitening for interpretable image recognition. *Nat. Mach. Intell.* **2**, 772–782. (doi:10.1038/s42256-020-00265-z)
101. Koh PW, Nguyen T, Tang YS, Mussmann S, Pierson E, Kim B, Liang P. 2020 Concept bottleneck models. In *Proc. of the 37th Int. Conf. on Machine Learning* (eds HD III, A Singh), vol. 119. Proc. of Machine Learning Research, pp. 5338–5348. New York, NY: PMLR.
102. Clough J, Oksuz I, Puyol-Antón E, Ruijsink B, King A, Schnabel J. 2019 In *Global local interpretability for cardiac MRI classification*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). *22nd Int. Conf. on Medical Image Computing and Computer-Assisted Intervention MICCAI 2019; Conference date: 13-10-2019 Through 17-10-2019*, pp. 656–664. Berlin, Germany: Springer (doi:10.1007/978-3-030-32251-9_72)
103. Graziani M, Andrearczyk V, Müller H. 2018 Regression concept vectors for bidirectional explanations in histopathology. In *Understanding and interpreting machine learning in medical image computing applications* (eds D Stoyanov *et al.*), pp. 124–132. Cham: Springer International Publishing.
104. Mincu D *et al.* 2021 *Concept-based model explanations for electronic health records*, pp. 36–46. New York, NY: Association for Computing Machinery.
105. Cai CJ *et al.* 2019 Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proc. of the 2019 CHI Conf. on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019*, pp 1–14. New York NY: ACM.
106. Tversky A. 1977 Features of similarity. *Psychol. Rev.* **84**, 327. (doi:10.1037/0033-295X.84.4.327)
107. Tenenbaum J, Griffiths T. 2001 Generalization, similarity, and Bayesian inference. *Behav. Brain Sci.* **24**, 629–640. (doi:10.1017/S0140525X01000061)
108. Carey S. 1985 *Conceptual change in childhood*. Cambridge, MA: MIT Press.
109. Carey S. 1991 Knowledge acquisition: enrichment or conceptual change. *The epigenesis of mind: essays on biology and cognition*, pp. 257–291. New York, NY: Psychology Press.
110. Piantadosi ST, Tenenbaum JB, Goodman ND. 2012 Bootstrapping in a language of thought: a formal model of numerical concept learning. *Cognition* **123**, 199–217. (doi:10.1016/j.cognition.2011.11.005)
111. Fyfe ER, McNeil NM, Son JY, Goldstone RL. 2014 Concreteness fading in mathematics and science instruction: a systematic review. *Educ. Psychol. Rev.* **26**, 9–25. (doi:10.1007/s10648-014-9249-3)
112. Pyatkin V, Hwang JD, Srikumar V, Lu X, Jiang L, Choi Y, Bhagavatula C. 2022 Reinforced clarification question generation with defeasibility rewards for disambiguating social and moral situations. (https://arxiv.org/abs/2212.10409)
113. Corbett A. 2021 Cognitive computer tutors: solving the two-sigma problem. In *Int. Conf. on User Modeling*, pp. 137–147. Berlin, Germany: Springer.
114. Pavlik Jr PI, Cen H, Koedinger KR. 2009 Performance factors analysis—a new alternative to knowledge tracing. *Online Submission*.
115. Piech C, Bassen J, Huang J, Ganguli S, Sahami M, Guibas LJ, Sohl-Dickstein J. 2015 Deep knowledge tracing. *Adv. Neural Inform. Process. Syst.* **28**, 505–513.
116. Kim B, Glassman E, Johnson B, Shah J. 2015 iBCM: interactive Bayesian case model empowering humans via intuitive interaction. *MIT CSAIL Technical Reports*. Boston, MA: CSAIL.

117. Rafferty AN, Brunskill E, Griffiths TL, Shafto P. 2011 Faster teaching by POMDP planning. In *Int. Conf. on Artificial Intelligence in Education*, pp. 280–287. Berlin, Germany: Springer.
118. Nie A, Brunskill E, Piech C. 2021 Play to grade: testing coding games as classifying Markov decision process. *Adv. Neural Inf. Process. Syst.* **34**, 1506–1518.
119. Malik A, Wu M, Vasavada V, Song J, Coots M, Mitchell J, Goodman N, Piech C. 2019 Generative grading: near human-level accuracy for automated feedback on richly structured problems. (https://arxiv.org/abs/1905.09916)
120. Tack A, Piech C. 2022 The AI teacher test: measuring the pedagogical ability of blender and GPT-3 in educational dialogues. (https://arxiv.org/abs/2205.07540)
121. Scholl B, Tremoulet P. 2000 Perceptual causality and animacy. *Trends Cogn. Sci.* **4**, 299–309. (doi:10.1016/S1364-6613(00)01506-0)
122. Konkle T, Caramazza A. 2013 Tripartite organization of the ventral stream by animacy and object size. *J. Neurosci.* **33**, 10 235–10 242. (doi:10.1523/JNEUROSCI.0983-13.2013)
123. Deen B, Koldewyn K, Kanwisher N, Saxe R. 2015 Functional organization of social perception and cognition in the superior temporal sulcus. *Cereb. Cortex* **25**, 4596–4609. (doi:10.1093/cercor/bhv111)
124. Richardson H, Lisandrelli G, Riobueno-Naylor A, Saxe R. 2018 Development of the social brain from age three to twelve years. *Nat. Commun.* **9**, 1027. (doi:10.1038/s41467-018-03399-2)
125. Spelke ES. 2022 *What babies know: core knowledge and composition*, **vol. 1**. Oxford, UK: Oxford University Press.
126. Fischer J, Mikhael JG, Tenenbaum JB, Kanwisher N. 2016 Functional neuroanatomy of intuitive physical inference. *Proc. Natl Acad. Sci. USA* **113**, E5072–E5081. (doi:10.1073/pnas.1610344113)
127. DiCarlo JJ, Zoccolan D, Rust NC. 2012 How does the brain solve visual object recognition?. *Neuron* **73**, 415–434. (doi:10.1016/j.neuron.2012.01.010)
128. Johnson S, Slaughter V, Carey S. 1998 Whose gaze will infants follow? Features that elicit gaze-following in 12-month-olds. *Dev. Sci.* **1**, 233–238. (doi:10.1111/1467-7687.00036)
129. Weisman K, Dweck CS, Markman EM. 2017 Rethinking people's conceptions of mental life. *Proc. Natl Acad. Sci. USA* **114**, 11 374–11 379. (doi:10.1073/pnas.1704347114)
130. Weisman K *et al.* 2021 Similarities and differences in concepts of mental life among adults and children in five cultures. *Nat. Hum. Behav.* **5**, 1358–1368. (doi:10.1038/s41562-021-01184-8)
131. Brink KA, Wellman HM. 2020 Robot teachers for children? Young children trust robots depending on their perceived accuracy and agency. *Dev. Psychol.* **56**, 1268. (doi:10.1037/dev0000884)
132. Flanagan T, Wong G, Kushnir T. In press. The minds of machines: children's beliefs about the experiences, thoughts, and morals of familiar interactive technologies. *Dev. Psychol.* **59**, 37036664. (doi:10.1037/dev0001524)
133. Dietz G, Outa J, Howe L, Landay J, Gweon H. *Proc. of the 44th Ann. Conf. of the Cog Sci Society*. Cognitive Science Society.
134. Carey S. 2000 The origin of concepts. *J. Cogn. Dev.* **1**, 37–41. (doi:10.1207/S15327647JCD0101N_3)
135. Gopnik A, Wellman HM. 1992 Why the child's theory of mind really is a theory. *Mind Lang.* **7**, 145–171. (doi:10.1111/j.1468-0017.1992.tb00202.x)
136. Hevelke A, Nida-Rümelin J. 2015 Responsibility for crashes of autonomous vehicles: an ethical analysis. *Sci. Eng. Ethics* **21**, 619–630. (doi:10.1007/s11948-014-9565-5)
137. Marchant GE, Lindor RA. 2012 The coming collision between autonomous vehicles and the liability system. *Santa Clara L. Rev.* **52**, 1321.
138. Hidalgo CA, Orghian D, Canals JA, De Almeida F, Martín N. 2021 *How humans judge machines*. New York, NY: MIT Press.
139. Awad E, Levine S, Kleiman-Weiner M, Dsouza S, Tenenbaum JB, Shariff A, Bonnefon JF, Rahwan I. 2020 Drivers are blamed more than their automated cars when both make mistakes. *Nat. Hum. Behav.* **4**, 134–143. (doi:10.1038/s41562-019-0762-8)
140. Jiang L *et al.* 2022 Can machines learn morality? the Delphi experiment. arXiv:2110.07574 [cs.CL], Apr 30.
141. Nguyen A, Yosinski J, Clune J. 2015 Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 427–436. Silver Spring, MD: IEEE.

142. Bear DM *et al.* 2021 Physion: evaluating physical prediction from vision in humans and machines. (https://arxiv.org/abs/2106.08261)
143. Firestone C. 2020 Performance vs. competence in human–machine comparisons. *Proc. Natl Acad. Sci. USA* **117**, 26 562–26 571. (doi:10.1073/pnas.1905334117)
144. Brockbank E, Wang H, Yang J, Mirchandani S, Bıyık E, Sadigh D, Fan JE. 2022 How do people incorporate advice from artificial agents when making physical judgments? (https://arxiv.org/abs/2205.11613)
145. Rahwan I *et al.* 2019 Machine behaviour. *Nature* **568**, 477–486. (doi:10.1038/s41586-019-1138-y)
146. Omidshafiei S, Kapishnikov A, Assogba Y, Dixon L, Kim B. 2022 Beyond rewards: a hierarchical perspective on offline multiagent behavioral analysis. (https://arxiv.org/abs/2206.09046)
147. Grupen N, Jaques N, Kim B, Omidshafiei S. Concept-based understanding of emergent multi-agent behavior. In *Deep Reinforcement Learning Workshop NeurIPS 2022*. New York, NY: Curran Associates, Inc.
148. Frank MC *et al*. 2023 *Experimentology: an open science approach to experimental psychology methods*. Boston, MA: MIT Press.
149. Shu T, Bhandwaldar A, Gan C, Smith K, Liu S, Gutfreund D, Spelke E, Tenenbaum J, Ullman T. 2021 Agent: a benchmark for core psychological reasoning. In *Int. Conf. on Machine Learning*, pp. 9614–9625. New York, NY: PMLR.
150. Gandhi K, Stojnic G, Lake BM, Dillon MR. 2021 Baby intuitions benchmark (BIB): discerning the goals, preferences, and actions of others. *Adv. Neural Inf. Process. Syst.* **34**, 9963–9976.
151. Doshi-Velez F, Kim B. 2017 Towards a rigorous science of interpretable machine learning. (https://arxiv.org/abs/1702.08608)