Trends in
Cognitive Sciences

CellPress

Imagining and building wise machines: the centrality of AI metacognition

Samuel G.B. Johnson [1,*], Amir-Hossein Karimi[2], Yoshua Bengio[3], Nick Chater[4], Tobias Gerstenberg[5], Kate Larson[6], Sydney Levine[7], Melanie Mitchell[8], Iyad Rahwan[9], Bernhard Schölkopf[10], and Igor Grossmann[1,*]

Although artificial intelligence (AI) has become increasingly smart, its wisdom has not kept pace. In this opinion article, we examine what is known about human wisdom and sketch a vision of its AI counterpart. We introduce human wisdom as strategies for solving intractable problems—those outside the scope of analytic techniques—including both 'object-level' strategies, such as heuristics (for managing problems), and 'metacognitive' strategies, such as intellectual humility, perspective-taking, or context adaptability (for managing object-level task fit). We argue that AI systems particularly struggle with this type of metacognition. Wise metacognition would lead to AI that is more robust to novel environments, explainable to users, cooperative with others, and safer by risking fewer misaligned goals with human users. We discuss how wise AI might be benchmarked, trained, and implemented.

Highlights

We examine the why and the how of building wise artificial intelligence.

Wisdom helps humans navigate intractable problems through object-level strategies (for managing problems) and metacognitive strategies (for managing object-level strategies).

Wise metacognition includes strategies such as intellectual humility, perspective-taking, and context adaptability.

Wise artificial intelligence, through such improved metacognitive strategies, would be more robust to new environments, explainable to users, cooperative in pursuing shared goals, and safe in avoiding both prosaic and catastrophic failures.

We suggest several approaches to benchmarking wisdom, training wise reasoning strategies, and adapting artificial intelligence architectures for metacognition.

Where does artificial intelligence still struggle?

Despite recent breakthroughs, artificial intelligence systems (AIs) still face critical shortcomings. They struggle in novel, unpredictable environments, lacking robustness. Their computations are opaque, creating a problem of explainability [1]. Their challenges with communication and credibility create barriers to cooperation [2]. These shortcomings limit our ability to harness the benefits of artificial intelligence (AI) while avoiding risks and ensuring safety [3]. As AIs increasingly act as agents in the world, these problems will be exacerbated.

But AIs are not the only intelligent agents that must solve these problems—we humans also face analogs of each of them. Might our own solutions yield some clues as to how AIs might do so more effectively?

We argue that one core set of capabilities underlies humans' ability to make robust decisions, explain their reasoning, achieve goals cooperatively, and interact safely—**wisdom** (see Glossary). We examine the function and mechanisms of human wisdom, concluding that wisdom serves to solve **intractable problems** and proceeds via a suite of complementary **object-level strategies** (which provide possible solutions to problems) and perspectival **metacognitive strategies** (which are necessary to decide among the solutions). We then consider how humans use these mechanisms to solve our versions of the robustness, explainability, cooperation, and safety problems. By analogy, we suggest that fostering wisdom in AIs—particularly wise metacognition—will help address these problems.

[1]Department of Psychology, University of Waterloo, Waterloo, Ontario, Canada
[2]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada
[3]Department of Computer Science and Operations Research, Université de Montréal, Montreal, Quebec, Canada
[4]Behavioural Science Group, Warwick Business School, Coventry, UK
[5]Department of Psychology, Stanford University, Stanford, CA, USA
[6]University of Waterloo, Cheriton School of Computer Science, Waterloo, Ontario, Canada
[7]Google DeepMind, London, UK
[8]Santa Fe Institute, Santa Fe, NM, USA

What is wisdom?

Consider these examples of human wisdom:

- Willa's children are bitterly arguing about money. Willa draws on her life experience to explain why they should, instead, compromise in the short term and prioritize their sibling relationship in the long term.
- Daphne is a world-class cardiologist. Nonetheless, she consults with a junior colleague when she recognizes that he knows more about a patient's history than she does.
- Ron is a political consultant who formulates possible scenarios to ensure his candidate will win. He not only imagines best-case scenarios but also imagines that his client has lost the election and considers what might have caused the loss.

Why do we intuit some abilities (applying life experience, being intellectually humble, reflective scenario planning) as 'wise' but not others (solving tricky integrals, cracking clever jokes, composing beautiful sonnets)? Accounts of wisdom highlight a wide array of characteristics [4–10] (Box 1). In our view, differences across theories mask important generalizations about wisdom's function and mechanisms [13].

The function of human wisdom: navigating intractable situations

If we lived in a textbook, we would not need wisdom. All problems would have correct answers, and the world would advertise the information required to find them. Natural selection would have made us nothing more or less than master statisticians, merciless optimizers, and lightning

[9]Max Planck Institute for Human Development, Berlin, Germany
[10]Max Planck Institute for Intelligent Systems, Tübingen, Germany

*Correspondence:
samuel.johnson@uwaterloo.ca
(S.G.B. Johnson) and
igrossma@uwaterloo.ca (I. Grossmann).

Box 1. Wisdom and metacognition

Though philosophers have debated wisdom for millennia, empirically grounded models are recent [4–10]. The Berlin Wisdom Model defines wisdom as expertise in important and difficult life matters, combining knowledge (e.g., about human nature) with certain metacognitive strategies that are sensitive to context, value pluralism, and uncertainty [6]. The MORE Model highlights how wise people build psychological resources—such as managing uncertainty and developing open reflectiveness toward experiences and perspectives—to cope with life's challenges [7]. Balance Theory emphasizes how wise people deploy their knowledge and skills toward the common good by balancing interests (theirs, others', and society's) and time horizons (short- and long-term) [10].

Emerging consensus models integrate these perspectives, either conceptually [8] or by surveying wisdom researchers directly [4]. Across approaches, wisdom converges on a cluster of metacognitive skills—context sensitivity, intellectual humility, interest balancing, and perspective integration—which we term perspectival metacognition. Rooted in philosophical perspectivism, it shifts the goal of reasoning from finding a single 'correct' answer toward achieving a state of maximal situational clarity, attained by evaluating and coordinating competing interpretations.

Although individuals vary in these skills, most people show them to some extent, for example, when planning ahead or coordinating within social groups [4]. This view challenges the notion that wisdom is reserved for a rare elite; instead, most humans exhibit moments of both wisdom and folly [9].

Not all metacognition is perspectival. Whereas some metacognitive strategies (e.g., monitoring memory and checking reasoning) optimize performance on well-structured tasks with clear accuracy criteria, perspectival metacognition specifically concerns multiple, often incommensurable, perspectives. Recruited for ill-structured, value-laden social problems, in which multiple, partly incompatible, standpoints must be coordinated rather than simply judged as right or wrong, this subset of metacognition moves beyond egocentric reasoning toward balancing interests, adapting to context, and recognizing epistemic limits when decisions affect others.

Although metacognition is central to wisdom, it does not exhaust it; most wisdom models also treat concern for others and the common good as central components [4,8]. One possibility is that, in many real-world contexts, such as repeated interactions with the same partners [11] or when one's reputation matters to third parties [12], the most effective way to deal with difficult life challenges, even from a self-interested standpoint, is to prioritize the common good.

calculators. Indeed, in some domains—such as low-level visual processing—we approximate this ideal.

Yet, social interaction and decision-making in an unstructured, ever-changing world require further tools [14]. Such problems are often intractable in one or more ways. This can happen because of ambiguities in goals—conflicting values that cannot be put on the same scale ('incommensurable' [15]) or a potential outcome changes underlying preferences ('transformative' [16]). It can happen because probabilities cannot be assigned to possible outcomes—the outcomes cannot be enumerated ('radically uncertain' [17]), there is a strong dependency on initial conditions ('chaotic' [18]), the underlying process is changing over time ('nonstationary'), or the situation is far beyond experience ('out of distribution'). It can also happen if the optimal outcome is calculable only with infeasible resources ('computationally explosive').

Our earlier examples of wisdom featured such intractability. Wisdom helped Willa understand how to make an incommensurable trade-off, helped Daphne navigate an out-of-distribution situation, and helped Ron make useful forecasts despite his ignorance about the radically uncertain future.

Mechanisms of human wisdom: metacognitive strategy selection

We argue that wisdom manages intractable problems by cultivating and deploying two types of strategies [4,13] (Figures 1 and 2): object-level strategies to manage the problem itself (i.e., the 'object' of judgment) and a cluster of metacognitive strategies to manage those object-level strategies, particularly when they conflict [19,20].

Object-level strategies yield candidate solutions to intractable problems. Many object-level strategies are **heuristics**—rules of thumb that rely on a small number of inputs and do not attempt a complex analysis [21] but may approximate it [22]. For example, Willa and Ron may have used heuristics such as 'prioritize family relationships' and 'avoid the worst-case scenario'. Heuristics often work well, despite requiring less computation than optimization, because they focus on just the most relevant information, reducing the chances of overfitting [21]. Much of 'folk wisdom' comprises culturally evolved heuristics transmitted across generations (e.g., deference to elders) [23].

The trouble with object-level strategies is their multiplicity. Heuristics can conflict ('look before you leap' vs. 'he who hesitates is lost'), and other classes of strategies, varying in computational complexity, coexist with heuristics. In **narrative thinking**, a reasoner uses causal knowledge and analogies to construct a mental model that can explain a situation, generate predictions, and evaluate choices [14,24], as when Ron draws on his knowledge and experiences to generate worst-case scenarios. These intuitive strategies also coexist with **decision technologies**, such as explicit analytic strategies, as when Daphne uses risk-scoring algorithms as one input to clinical decision-making. Wisdom requires us not just to have these strategies, but to manage them effectively.

Even a well-tailored suite of object-level strategies falls short of wisdom. First, even simple strategies depend on information; an **input-seeking process** is required. (Ron must check if he has the relevant facts for his scenarios and fill any gaps.) Second, strategies often yield conflicting advice; a **conflict resolution process** is required to select the best strategy for each situation [25]. (Should Daphne follow the strategy 'trust your judgment' or 'trust knowledge-able experts'?) Third, strategies can break under unfavorable conditions, as when the underlying pattern changes unpredictably; an **outcome-monitoring process** is required to safeguard

Glossary

AI alignment: ensuring that AIs pursue the goals intended by ('aligned with') their human users.

Benchmark: a set of standard tasks on which AIs can be compared with one another and to humans for a given capacity.

Commitment: the ability to make a credible promise that will be kept at a later time, particularly as a means of incentivizing a mutually beneficial cooperative agreement.

Context window: the sliding window of text that a GenAI model has access to (can 'remember') when formulating its output.

Conflict resolution process: a type of metacognitive process that selects the best strategy when object-level strategies conflict.

Cooperative AI: AI that is able to pursue shared goals—with other AIs or with human users—through abilities including social understanding, communication, and credible commitment.

Decision technologies: organized procedures for making decisions, such as formal calculation.

Explainable AI: AI that can be effectively understood by users, for instance because the AI can effectively communicate its decisions and reasoning to users.

Heuristic: an object-level strategy that produces a solution to a problem without conducting a full analysis, typically by using a subset of the available information.

Input-seeking process: a type of metacognitive process that seeks the inputs required for object-level strategies to work.

Intractable problem: a problem that does not lend itself to analytic techniques, such as optimization.

Metacognitive strategy: a strategy that is used to manage other (especially object-level) strategies, including by seeking the required inputs, resolving conflicts among strategies, and monitoring the plausibility of outcomes.

Narrative thinking: an object-level strategy in which an individual constructs a causal and analogical model of a situation in order to understand a situation, predict how it will unfold, and evaluate potential choices.

Object-level strategy: a strategy that is used to produce a potential solution to a specific problem or task, such as a

Key figure

The relationship between object-level and metacognitive strategies in wise reasoning
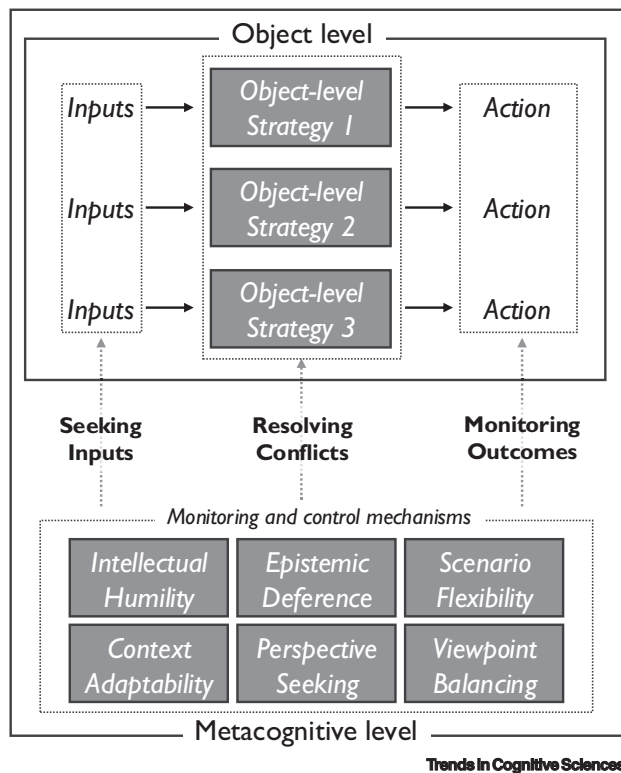
Figure 1. Object-level strategies (e.g., heuristics, narratives, and decision technologies) provide candidate actions for a given situation. Metacognitive monitoring and control processes regulate these strategies in three ways: obtaining the appropriate inputs, deciding which strategy to use when they conflict, and monitoring their outcomes to avoid catastrophic actions.

heuristic, narrative, or analytic procedure.

Outcome-monitoring process: a type of metacognitive process that checks whether the outcomes of the selected object-level strategy are plausible (also called 'sanity checking').

Perspectival metacognition: a subset of metacognitive skills for managing and integrating perspectives on a situation.

Robust AI: AI that works effectively in novel environments because it is reliable (similar inputs yield similar outputs), unbiased (not systematically mistaken), and flexible (able to generalize to novel inputs).

Safe AI: AI that avoids risks associated with harmful failures, which can include both incompetence (e.g., errors because the AI is not robust) or malevolence (e.g., malfeasance because the AI is not aligned).

Wisdom: a suite of abilities used to solve intractable problems, comprising both metacognitive strategies (e.g., intellectual humility) and object-level strategies (e.g., culturally transmitted heuristics).

against nonsensical outcomes. (Willa would question her usual advice if one child was taking advantage of the other.)

Navigating this complexity requires the ability to monitor and adapt object-level strategies [26–28] using **perspectival metacognition** [4]—strategies for coordinating perspectives, including one's own and others' (Box 1). Some are primarily epistemic: intellectual humility comprises awareness of what one does (or does not) know [29]; scenario flexibility involves considering the diverse ways that a scenario could unfold; context adaptability identifies features of a situation that make it comparable with or distinct from other situations [6]. Others have a social dimension: epistemic deference is a willingness to defer to others' expertise [30]; perspective seeking draws on multiple perspectives [6]; viewpoint balancing recognizes and integrates discrepant interests [10].

Perspectival metacognition contributes to the input-seeking, conflict resolution, and outcome-monitoring required to manage object-level strategies (Figure 1, Key figure). For example, perspective seeking is important for gaining relevant inputs, context adaptability is crucial for
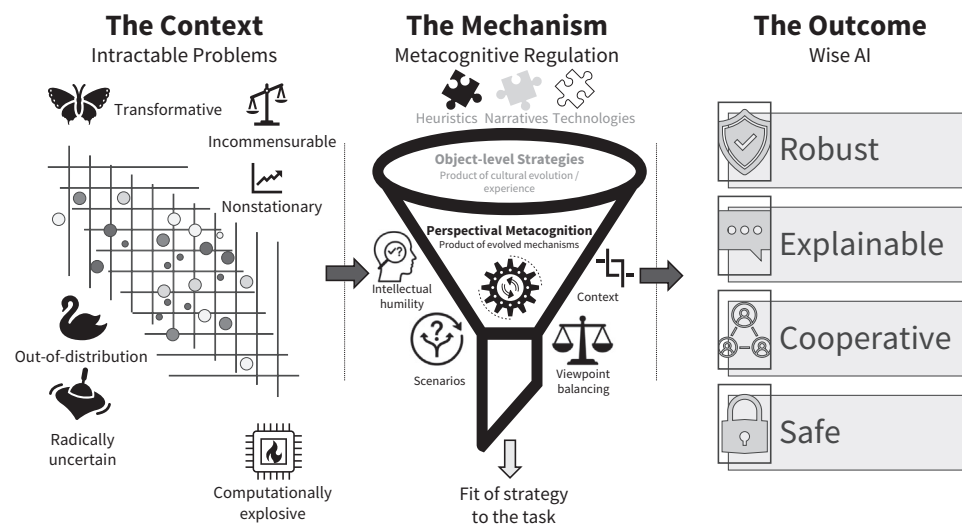
Figure 2. From intractability problems to wise AI. Left panel: wisdom functions specifically to solve intractable problems —situations that resist analytic optimization due to incommensurable values, radical uncertainty, or nonstationary environments. Center panel: to address these challenges, agents deploy object-level strategies (e.g., intuitive strategies, such as heuristics and narratives, as well as technologies, such as formal procedures) derived from experience or cultural evolution. However, these strategies are often insufficient because they conflict or lack necessary inputs. Perspectival meta-cognition serves as the regulatory control, applying epistemic strategies (e.g., intellectual humility: awareness of what one does and does not know; acknowledgment of uncertainty and one's fallibility) and social strategies (e.g., viewpoint balancing: recognizing and integrating discrepant interests) to select and adapt the correct object-level approach to the task at hand. Right panel: implementing this metacognitive architecture enables AI to move beyond a narrow view of intelligence as optimization toward wisdom, resulting in systems that are more robust, explainable, cooperative, and safe.

resolving conflicts among strategies in a context-sensitive way, and viewpoint balancing is one component of outcome monitoring. Often, these strategies work together. Daphne exhibits intellectual humility when she recognizes that she does not understand her patient's symptoms (recognizing that her existing object-level strategies are inappropriate); perspective seeking when she calls upon her colleague's expertise (seeking out new object-level strategies); context adaptability when she considers whether her patient's unique situation limits the relevance of her colleague's expertise (assessing the relevance of new object-level strategies); and, ultimately, epistemic deference when she adopts her colleague's view (accepting the outcome of the newly proposed strategy).

Toward wise AI: machine metacognition

Object-level strategies define the search space, whereas perspectival metacognition provides the tools for its wise navigation (Figure 2). Here, we focus on the latter because it has been the subject of comparatively less research (but see Box 2 regarding recent advances in AI metacognition broadly). We believe that perspectival metacognition is, therefore, the current weak link for wise AI. Here, we focus on generative artificial intelligence (GenAI) systems such as large language models (LLMs), but the arguments extend to other AI paradigms.

GenAI models do have rudimentary forms of metacognition [58]. They can monitor and control some of their neural activations in 'neurofeedback' paradigms [59]. They can classify math problems by solving procedure [60] and, on easier problems, can assess whether a step taken led in the correct direction [61]. They perform well on tests of situational competence for relatively unambiguous situations [62], and some models can use an inference-time search to decide when to stop searching.

Box 2. Metacognition in LLMs

Building on earlier precedents [31–34], interest in AI metacognition has exploded alongside LLMs [35].

Most work on LLM metacognition uses prompting. In chain-of-thought prompting, the model produces intermediate reasoning steps, which are added to the model's context window to inform subsequent processing [36]. This is metacognitive in that the model is asked to interpret its reasoning processes and use that interpretation to improve its reasoning. Extensions to chain-of-thought prompting require more complex metacognitive control, such as chains that backtrack or branch ('tree of thought' [37] and 'meta chain-of-thought' [38]), break problems into sequences of easier to harder sub-problems ('least-to-most prompting' [39] and 'plan-and-solve prompting' [40]), or consider multiple possible reasoning chains ('self-consistency' [41]). Other prompting techniques more explicitly incorporate wisdom-related metacognition. For example, metacognitive prompting [42] asks LLMs to consider several metacognitive queries at the same time as a primary query. Similarly, prompting a model to consider, 'could you be wrong?' leads it to consider data in its training set highlighting potential errors and biases [43].

Other approaches alter the model architecture by introducing additional modules or agents. The most common is a module for evaluating or monitoring proposed outputs. For example, EXAR uses a 'meta-validator' module that assesses the outputs of a model and uses those assessments to fine-tune the model [44]. MIRROR separates a 'thinker' that creates an inner monologue and a 'talker' that uses the inner monologue as context for communicating with the user [45]. Another common addition is an explicit module or reasoning step for evaluating model capabilities, learning progress (e.g., in MAGELLAN [46]), or the probability of task success (e.g., in MUSE [47]).

Still other techniques include mapping and evaluating the intermediate steps in chain-of-thought reasoning in a latent space [48], monitoring hidden layers for anomalies [49], comparing neural activation patterns in novel cases to baselines with known metacognition [50], treating metacognitive tasks, such as self-awareness and self-evaluation, as Bayesian learning problems [51], and incorporating 'fast' and 'slow' reasoning modes for routine versus complex problems [52].

Models with enhanced metacognition have been applied to a growing array of problems—few-shot classification [53], external tool selection [50], improving the use of external knowledge sources [54], inferring beliefs and desires [55], accurately classifying linguistic nuance, such as sarcasm [56], and navigating simulated scenarios, such as a zombie apocalypse [57]. Yet, these techniques remain imperfect and not all are wisdom related (Figure 2).

At the same time, they struggle with other metacognitively loaded tasks [58]. They often 'hallucinate' an answer rather than admit ignorance [63], and they struggle to understand their goals [64], capabilities [64], and the strength of their evidence [65]. This cluster of epistemic failings has been argued to be symptomatic of a broader 'metacognitive myopia' [66].

What are the potential benefits of wise AI?
We suggest that deficits in AI wisdom—and perspectival metacognition in particular—underlie failures in robustness, explainability, cooperation, and safety (Figure 2).

Robustness
Given the range of intractable environments in which intelligent systems must operate, three failures of robustness are common: a system can be unreliable—given similar inputs, a system can produce wildly different outputs. This can be caused by applying different strategies each time or by applying a strategy that produces inconsistent results. A system can be biased—the output is systematically wrong or nonrepresentative in a predictable direction. Or a system can be inflexible—novel inputs lead to lower-quality outputs.

Human wisdom combines object-level and metacognitive strategies to adapt robustly across environments. Object-level strategies such as heuristics can be beneficial because they sometimes outperform analytic optimization by avoiding data overfitting [21], especially in novel, out-of-distribution contexts (but see [67]). These strategies are supported by wise metacognition, which helps reasoners learn new information from other perspectives and discern its relevance, balance the competing urges to simplify and optimize, and avoid catastrophic error by checking the plausibility of a strategy's output.

For similar reasons, wise AI would be more **robust AI** in all three senses. It would be more reliable; its monitoring processes would evaluate whether it is sensible to use different strategies in comparable situations and reject excessively inconsistent strategies. It would also be less biased; since biased outputs usually result from biased inputs, a wise AI would reflect on its training data or models of the world, to identify sample deficiencies in its training data (perhaps requesting additional data) and to understand the causal processes by which biases resulted (correcting for those biases). Finally, wise AI would be more flexible; it would moderate its confidence in novel situations and would reduce, manage, and navigate uncertainty.

Explainability
Opaque AI can produce puzzling outputs, difficult-to-diagnose errors, and barriers to collaboration [1]. Although cognitive scientists disagree about the extent of introspective access in humans [68], all theories agree that metacognition is necessary for justifying decisions to ourselves and others. Thus, wise AI would likely be more explainable.

One possibility is that, in humans, consciously accessible metacognitive strategies guide behavior. When we report our thought processes, we are reporting observations. For instance, the decision to moderate confidence in a prediction could be caused by a conscious recognition of ignorance, which can then be reported. The explainability problem is then reduced to selecting which of those observations to report, that is, which are the most relevant causal antecedents of the output.

On the opposite extreme, the mind may be 'flat' [69]—it does not contain hidden depths of reasons that can be uncovered through introspection. When we report our thought processes, we report inferences ('stories'), not observations. The reasoner observes the outputs of her strategies and reasons backward to their possible causes [70]. These inferences may often be incorrect [71], yet they are often useful justifications that, when expressed, constrain future thought and behavior. Since metacognition itself is not observable but only inferable, **explainable AI** would need to generate a useful narrative to make sense of its own actions—itself a metacognitive process.

Recent work suggests that even when using techniques such as chain-of-thought [72] or metacognitive prompting [73] (Box 2), models confabulate insight rather than genuinely introspect, generating explanations that are not 'faithful' to their underlying reasoning. Under the classical view, we would hope that techniques for improving the introspective accuracy of metacognition would yield more faithful explanations; if the mental flatness view is correct, all we can hope for is more useful post hoc reconstructions.

Cooperation
AIs increasingly behave within larger networks, requiring both AI–AI cooperation (e.g., autonomous vehicles negotiating traffic) and AI–human cooperation (e.g., surgical robots), while also influencing human–human cooperation (e.g., social media content curation). **Cooperative AI** [2,74] examines how AI can benefit all parties involved by navigating barriers to understanding, communication, and **commitment**. Wise object-level and metacognitive strategies are critical to how humans solve these problems, suggesting the same may be true for AI.

Cooperation requires understanding the social dynamics of the situation, including the likely actions taken by others. Since those actions depend on the beliefs and goals of agents, social understanding requires theory of mind [75], including the tacit ability to form joint plans to coordinate behavior [76]. In humans, this is accomplished through object-level strategies such

as first-person simulation (putting oneself in the other's shoes) [77] and third-person, theory-based reasoning (e.g., assuming that the agent is rational [78]).

Cooperation depends equally on communication—selecting and sending information to potential partners. Incoming information must be filtered to act on what is useful and ignore what is misleading or irrelevant [79]. Even young children develop object-level strategies for evaluating sources—tracking cues such as accurate past testimony and conflicts of interest [80]—and more sophisticated reasoners can check whether the reasoning itself is valid [81]. Such 'epistemic vigilance' mechanisms make credible communication among humans possible; without a means of assessing communication, the risk of exploitation would undermine trust.

Cooperation can unravel when long-term incentives diverge, so humans have evolved ways to make credible commitments. Third-party social judgments—introducing potential punishment and reputational risk—impose external costs on defection [82], while emotions such as shame and guilt impose internal costs [83]. Humans sharing a cultural and psychological context can assume these costs as common ground, promoting credible commitment.

Wise metacognition is required to manage these object-level mechanisms effectively [84,85], resolving conflicts among strategies (e.g., when accuracy cues diverge), assessing their appropriateness (e.g., whether one can evaluate a chain of argumentation), and seeking appropriate inputs (e.g., knowing the capabilities of the other counterparty). This last point is particularly important for cooperative AI, which could overestimate the abilities of humans or lack common ground, such as a shared emotional system.

Safety

Concerns about AI safety span the prosaic to the cataclysmic [3,86]. For now, the main safety risks are simply that systems we come to rely on fail us: a shoddy surgical robot, incompetent tax advice, or a biased parole algorithm. Machine metacognition can help avoid such failures [87]. AIs with well-calibrated confidence can target the most likely risks; appropriate self-models would help AIs anticipate failures; and continual monitoring of their performance would facilitate recognition of high-risk moments.

Some worry, however, that in the future, superintelligent machines will pose an existential risk to humanity if their goals are not 'aligned' with ours [88]. This concern arises from two observations: (i) predefined goals are likely to be mis-specified or become obsolete, and (ii) a powerful AI could be difficult to curtail if it aggressively pursued the wrong goals. Bostrom [88] illustrates both points in his parable of the paper clip–maximizing AI that converts the Earth into paper clips and kills all humans in its way.

The goal of **AI alignment** [3] is to prevent such mismatches between the goals of an AI and its users—an exceedingly difficult task due to the many assumptions that are unspoken and potentially unshared. Wisdom is crucial to navigating such problems—first, because goal specification is a prototypical example of an intractable problem for which we deploy wisdom; and second, because humans rely on 'common sense' wisdom to fill in such unspoken assumptions and make tacit agreements [89].

Indeed, we suspect that engineering wise social interaction—in addition to or perhaps instead of alignment—may be necessary to achieve alignment's goals. Alignment faces not only technical problems but also conceptual ones. Who should we align AI to? People differ in their goals (e.g., believing GenAI should solely aim to provide accurate information versus avoiding the

reinforcement of harmful stereotypes) and values (e.g., crosscultural and religious differences in maximizing happiness vs. liberty) [90]. Should we increase the average human well-being, its sum, or care for the whole biosphere? And why assume that today's values are the right ones, given profound shifts even over recent history [91]? Aligning AI to current values would risk reifying those values as 'the right' values, stalling future social progress.

A two-pronged, wisdom-oriented approach to **safe AI** may be more promising.

First, AIs must themselves implement wise reasoning—aligning them with the right object-level and metacognitive strategies rather than the 'right' values. For example, one object-level strategy may be a bias toward inaction (not executing an action if it risks harm according to one of several possibly conflicting human norms), which in turn requires metacognitive regulation (learning what those conflicting perspectives are and avoiding overconfidence).

Second, we must consider how AIs fit into a broader institutional ecosystem. Institutions such as governments and markets address the 'alignment' problem that we humans have, ideally channeling our discrepant interests and values into socially productive directions. It is useful to think of AI not merely as an external tool influencing society but as a new type of agent within society, embedded in pairwise interactions and, increasingly, broader institutions. If channeled effectively through institutions, metacognitively wise AI can enhance social evolution rather than undermine it. Both human and artificial agents in society should continue to allow our values to evolve toward a shared reflective equilibrium [92], bringing situation-specific judgments and general moral principles into alignment with one another through iterative adjustments.

How might we build wise AI?

Before considering how we can build wise AI, first consider how nature has built us to be wise. We suspect that metacognitive abilities are primarily evolutionary adaptations built into the architecture of the human brain, being fundamental across any context, whereas object-level strategies are primarily acquired through experience, including socialization and didactic learning, due to their great situational variability. While acknowledging the role of development, culture, and self-reflection for metacognition [93,94], and biological evolution for object-level strategies [95], we take 'object-level = development' and 'metacognition = evolution' as a starting point. If so, this suggests that implementing object-level and metacognitive wisdom may require different strategies.

In humans, object-level strategies such as heuristics are typically acquired through trial and error and social learning. Since wise heuristics are often domain specific, exhaustively specifying these rules is likely doomed for the same reasons that rule-based expert systems in AI failed. Instead, allowing AI systems to learn from experience [96] and from others [97] may be more promising.

The analogy to the human case suggests, however, that experience alone is unlikely to suffice for training metacognition. One problem is that in typical AI training, a loss function is minimized, which is defined over the model's outputs rather than its reasoning. Although this may indirectly select for sound decision-making strategies, the poor explainability of many state-of-the-art models makes it difficult to determine what those strategies are; an output may please a human judge for the wrong reasons. Such a system would often emulate the decisions of a wise human, but it would not itself be metacognitively wise.

How might one get around this problem? Optimistically, standard LLM training techniques could be modified. For example, a two-step training process could be implemented in which a model is first trained for wise strategy selection directly (e.g., correctly identifying when to be intellectually

humble) and then trained to use those strategies correctly (e.g., carrying out intellectually humble behavior). Alternatively, one could present models with **benchmark** cases, request them to produce both their metacognitive strategy and their output, and then reward only the correct combination of strategy and output [98]. In either case, models could be trained against what a wise human would do or against the acceptability of its explanations for its choices.

Perhaps, however, no amount of training will get current models to human-level metacognition, just as no amount of language exposure will get a squirrel to talk. On this view, the 'innate' architecture of current models is not up to the task. LLMs work by generating the next token (i.e., word or word part) based on the input in their **context window**. At first, this input comprises the user's prompt; after the model is run to generate the first token in its response, this token is added to the context window, and the model is re-run to generate the second response token, and so on. This process does not involve feedback from later layers to earlier ones, and it is backward-looking—it predicts one word ahead based on its input and output so far, rather than explicitly planning ahead. This process can yield surprisingly intelligent outputs—and even some degree of planning (e.g., rhyming in a poem [99])—given enough parameters and data. Yet, given their lack of explicit planning, perhaps it is unsurprising that LLMs struggle with meta-cognition, which requires reflecting on one's thoughts and devising strategies to regulate them. Changes to model prompting and architecture may be required, not just changes to training. Box 2 describes some ongoing efforts in this spirit, while Table 1 lists some more speculative ideas.

Evaluating machine wisdom

Once we build a wise machine, how will we know it? Wisdom is context sensitive, so a bench-mark input must contain sufficient detail to match the rich context of a real-world situation. More-over, since wisdom is about the reasoning underlying strategy selection, any evaluation procedure must judge not only the outcome but also the precipitating process.

Existing benchmarking work in metacognition has focused on the calibration of confidence judg-ments [58,109]. An advantage of this narrower domain is that it is much more tractable than the perspectival metacognitive strategies we have discussed here, with well-developed methods that even work in nonhuman animals [110], lend themselves to computational modeling [111], and are able to separate performance on the cognitive versus metacognitive components of a task [109, 112]. Nonetheless, these tasks are domain specific, often constrained to well-defined laboratory environments, and do not yet capture the richness of everyday intractable problems that wise judgment handles.

To make progress, let us consider how other rich, complex constructs have been benchmarked. One approach is to collect tasks from psychology experiments, akin to benchmarking theory of mind or analogical reasoning [113,114]. Since these tasks are discussed in the literature (and appear in training data), the content must be replaced with structurally similar but superficially different problems [115,116]. However, since these tasks usually measure outcomes only and provide little context, this approach cannot be adopted wholesale for wisdom. An alternative approach—used to benchmark explanatory abilities [117]—is for domain experts to subjectively evaluate the quality of the model's outputs. This approach is well suited for evaluating reasoning (rather than outcomes) but requires some form of quantification to compare models.

One way to evaluate AI wisdom would start with tasks that measure wise reasoning in humans [118]. These tasks present participants with a social dilemma or a choice between seemingly incommensurable options and ask them to reflect on the next steps, with reflections scored on

Table 1. Engineering wiser AI via metacognition

Conceptual idea	Possible implementations
1. Explicit metacognitive checkpoints and error detection loops Integrate explicit reflective checkpoints into AI decision-making processes, forcing the AI to periodically evaluate coherence, reliability, and confidence in its reasoning. Implement continuous error detection loops, where an AI system revises internal strategies upon encountering prediction failures or contradictions	Introduce specific computational modules at defined decision points (e.g., transformer layers in LLMs) that assess output uncertainty (e.g., entropy and calibration error) and coherence metrics (e.g., consistency with past outputs). Implement error detection using confidence thresholds learned from validation data. For instance, pause execution to reassess decisions whenever model confidence falls below calibrated uncertainty thresholds, forcing conditional re-generation or seeking external verification
2. Epistemic source tagging and reliability updating Implement structured metadata that explicitly encodes epistemic reliability for training data sources. Allow systems to dynamically update their trust in data sources (provenance and lineage) based on the consistency of predictions and feedback, akin to human epistemic vigilance mechanisms	Precompute and embed metadata vectors, capturing reliability indicators (e.g., historical accuracy, domain expertise scores, publication credibility metrics), alongside raw tokens or data points. Train AI systems to dynamically adjust reliability scores using a simple online Bayesian updating mechanism: sources whose information frequently results in erroneous outputs or internal contradictions receive lowered reliability scores, reducing their influence during inference
3. Hierarchical and reflective reasoning architectures Employ hierarchical architectures inspired by cognitive models (e.g., ACT-R [100], Soar [101]), where a metacognitive layer explicitly monitors and selects object-level strategies. Develop explicit reflective sub-systems designed to audit internal consistency and logical coherence of reasoning outputs, promoting effective 'sanity checking'	Implement cognitive-architecture-inspired hierarchical models using explicit controller modules (meta-policy networks) to govern lower-level task-specific modules: (i) hybrid symbolic/subsymbolic approaches (e.g., OpenCog Hyperon [102], ACT-R style modules); (ii) reinforcement learning hierarchical controllers (e.g., FeUdal networks [103]). Introduce standalone 'auditor' modules explicitly trained to critique primary outputs for internal consistency, logical coherence, or sensitivity to constraints. For instance, chain-of-thought prompting [36] or future advanced reasoning modules explicitly trained as reasoning auditors
4. Transparency via metacognitive narration Design systems capable of transparently narrating their internal metacognitive reasoning ('thinking aloud' protocols) to users, aiding explainability and making reasoning easier to audit and debug	'Thinking Aloud' protocols: implement explicit model training on explanatory datasets or devise new chain-of-thought approaches that generate explicit narration of metacognitive reasoning steps in understandable language. Interactive debugging and auditing interfaces: build interactive visualization tools displaying model uncertainty, reasoning trails, or decision checkpoints to users or system auditors
5. Distributed and social metacognition Leverage multiagent reasoning and collective decision-making, analogous to human reliance on socially distributed cognition. Implement epistemic crosschecking and adversarial debate between multiple AI systems to mitigate individual AI overconfidence and misinformation propagation	Multiagent epistemic vigilance: multiple independent AI agents work collaboratively, requiring agreement or consensus for outputs on critical tasks. Concrete architectures: multiagent RL (MARL) [104], DAO-inspired decision-making [105]. Debate-based metacognitive crosschecking: AI reasoning outputs must pass adversarial debates or crossexaminations conducted by independently trained AI debaters before being finalized. Example frameworks: include OpenAI's debate-style AI safety approach [106] and Anthropic's Constitutional AI approach [107]
6. Scheduled offline replay and consolidation Use offline periods for AI systems to consolidate and 'reflect' on prior model runs, akin to one possible function of the human default mode network [94,108]	Allocate compute to periods during which outward action pauses while the model regenerates latent trajectories, pits alternative chains of thought against each other (self-consistency/debate), and refreshes its calibration curves before the next online cycle

ACT-R: Adaptive Control of Thought—Rational; DAO: decentralized autonomous organizations; RL: Reinforcement Learning.

prespecified criteria by human raters, such as experts. Novel and detailed variants of such scenarios could be presented to AIs, with their performance scored by either human raters or by other models (if their scores converge) [119]. It would be important to include problems that agentic AIs might confront in the future (e.g., whether to execute a debatably ethical request) to ensure they can reason wisely not only about humans but also about themselves.

Ultimately, the wisdom of increasingly autonomous AIs will be judged by human users and stakeholders. Prior benchmarking is a crucial start, but there is no substitute for interacting with the real world. Given this intrinsic limit on our ability to evaluate wisdom *ex ante*, this integration with the world must proceed slowly to minimize risks.

Concluding remarks

Building smarter machines comes with risks: AI with advanced capabilities might pursue undesirable goals. Is there a parallel concern about the unintended consequences of building wiser machines?

Perhaps not. Empirically, humans with wise metacognition show greater orientation toward the common good, including cooperation and responsiveness to others [84] (Box 1). Perhaps wise AI would have these qualities too.

Yet, an important ambiguity arises here about which we can only speculate: although current AI may not be wise, what shape would a future AI's wisdom take?

One possibility is that AI and human wisdom might sharply diverge. Human metacognition serves largely to economize scarce cognitive resources [120,121], and many biases may be side effects of solving this constrained optimization problem [122,123]. Given the more abundant computational resources of wise AI, this optimization problem may look very different from humans', AIs might rationally invest far more effort. Conversely, humans outsource much of their cognition to the social environment (as in the division of physical or cognitive labor [124,125]), including knowledge-generating institutions that are ever evolving. Distributed cognition of this sort is not yet a dominant paradigm in AI, and it is unclear what its (dis)advantages are compared with an extensive, integrated knowledge base.

Conversely, perhaps AI wisdom would converge considerably with human wisdom. AI wisdom also faces computational constraints, since computing can be costly. Moreover, heuristics work for AI for the same reasons as they work for humans: when we lack complete information, heuristics can perform well by implementing sensible, robust defaults. Finally, AIs may come to join our social environment—and perhaps reap some of the same social cognitive advantages as humans—as AI is increasingly integrated into human institutions [126].

Given these considerations, uncertainty remains (see Outstanding questions). What if we tried and failed to build wise AI? What if the characteristics of wise AI differ from those of a wise human, to the detriment of humans? To these concerns, we have three responses.

First, if the alternative were halting all AI progress, building wise AI would introduce added risks. But compared with the status quo—advancing capabilities at a breakneck pace without wise metacognition—the attempt to make machines intellectually humble, context adaptable, and adept at balancing viewpoints seems clearly preferable.

Second, at least in the medium term, AI will not act autonomously but will remain a collaborative tool to be used by and for humans, supporting rather than replacing human wisdom. In this

Outstanding questions

How might wise artificial intelligence (AI) inform—and be informed by—the cognitive science of human wisdom? For instance, how can computational modeling of human wisdom (including object-level and metacognitive strategies) and efforts to engineer machine wisdom be mutually enlightening?

What is the best approach to formalizing wise reasoning in mathematical approaches to AI robustness, explainability, cooperation, and safety?

How might the give-and-take of conversation between humans and AI lead to a form of shared wisdom? How should this potential for collaborative metacognition inform the design of AI systems?

Might AI wisdom exceed human wisdom? If so, how would we humans know?

How would the mass adoption of wise AI impact society? For example, could this lead to the offloading of metacognitive labor, resulting in a decline in human wisdom? Or could wise AI act as a cognitive prosthetic to enhance human wisdom in practice?

Could wise AI be subverted to malicious ends? Might wiser AI counter this problem or exacerbate it?

What can we learn about existential AI risks by studying wise and unwise human decision-making and institutional design around other existential risks, such as nuclear weapons?

Where would AI not benefit from wise metacognition—for instance, because the benefits are marginal relative to economic, environmental, or computational costs?

How would metacognitive AI systems scale up? How would the further integration of wise AI into human institutions impact the functioning of those institutions and AI itself?

What further considerations would be required to embody metacognition in robots?

sense, understanding how humans and AIs might work together to produce wise or foolish decisions becomes a crucial research agenda.

Finally, the qualities of robust, explainable, cooperative, and safe AI will amplify one another. Robustness facilitates cooperation (improving confidence from counterparties) and safety (avoiding failures in novel environments). Explainability facilitates robustness (aiding human intervention through transparency) and cooperation (enabling more effective communication). Cooperation facilitates explainability (accurate theory of mind about users) and safety (implementing shared values). Wise metacognition can lead to a virtuous cycle in AI, just as it does in humans. We may not know precisely what form wise AI will take, but it must surely be preferable to folly.

Declaration of interests

The authors declare no competing interests.

References

1. Dwivedi, R. *et al.* (2023) Explainable AI (XAI): core ideas, techniques, and solutions. *ACM Comput. Surv.* 55, 1–33
2. Dafoe, A. *et al.* (2020) Open problems in cooperative AI. *arXiv Prepr.* https://doi.org/10.48550/arXiv.2012.08630 *arXiv: 2012.08630*
3. Ji, J. *et al.* (2023) AI alignment: a comprehensive survey. *arXiv Prepr.* https://doi.org/10.48550/arXiv.2310.19852 *arXiv: 2310.19852*
4. Grossmann, I. *et al.* (2020) The science of wisdom in a polarized world: knowns and unknowns. *Psychol. Inq.* 31, 103–133
5. Ardelt, M. (2004) Wisdom as expert knowledge system: a critical review of a contemporary operationalization of an ancient concept. *Hum. Dev.* 47, 257–287
6. Baltes, P.B. and Smith, J. (2008) The fascination of wisdom: its nature, ontogeny, and function. *Perspect. Psychol. Sci.* 3, 56–64
7. Glück, J. and Bluck, S. (2013) The MORE Life Experience Model: a theory of the development of personal wisdom. In *The Scientific Study of Personal Wisdom* (Ferrari, M. and Weststrate, N.M., eds), pp. 75–98, Springer
8. Glück, J. and Weststrate, N.M. (2022) The wisdom researchers and the elephant: an integrative model of wise behavior. *Personal. Soc. Psychol. Rev.* 26, 342–374
9. Grossmann, I. (2017) Wisdom in context. *Perspect. Psychol. Sci.* 12, 233–257
10. Sternberg, R.J. (1998) A balance theory of wisdom. *Rev. Gen. Psychol.* 2, 347–365
11. Friedman, J.W. (1971) A non-cooperative equilibrium for supergames. *Rev. Econ. Stud.* 38, 1–12
12. Nowak, M.A. and Sigmund, K. (2005) Evolution of indirect reciprocity. *Nature* 437, 1291–1298
13. Grossmann, I. and Johnson, S.G.B. (2025) Cultivating wisdom through metacognition: a new frontier in decision-making under radical uncertainty. *J. Appl. Res. Mem. Cogn.* 14, 301–318
14. Johnson, S.G.B. *et al.* (2023) Conviction narrative theory: a theory of choice under radical uncertainty. *Behav. Brain Sci.* 46, e82
15. Walasek, L. and Brown, G.D. (2024) Incomparability and incommensurability in choice: no common currency of value? *Perspect. Psychol. Sci.* 19, 1011–1030
16. Paul, L.A. (2013) *Transformative Experience*, Oxford University Press
17. Kay, J. and King, M. (2020) *Radical Uncertainty: Decision-making beyond the Numbers*, Norton
18. Lorenz, E. (1993) *The Essence of Chaos*, University of Washington Press
19. Nelson, T.O. and Narens, L. (1990) Metamemory: a theoretical framework and new findings. In *Psychology of Learning and Motivation* (26) (Bower, G.H., ed.), pp. 125–173, Academic Press
20. Karlan, B. and Allen, C. (2024) Engineered wisdom for learning machines. *J. Exp. Theor. Artif. Intell.* 36, 257–272
21. Todd, P.M. and Gigerenzer, G. (2012) *Ecological Rationality: Intelligence in the World*, Oxford University Press
22. Parpart, P. *et al.* (2018) Heuristics as Bayesian inference under extreme priors. *Cogn. Psychol.* 102, 127–144
23. Edmondson, R. and Woerner, M.H. (2019) Sociocultural foundations of wisdom. In *The Cambridge Handbook of Wisdom* (Sternberg, R.J. and Glück, J., eds), pp. 40–68, Cambridge University Press
24. Glück, J. *et al.* (2005) The wisdom of experience: autobiographical narratives across adulthood. *Int. J. Behav. Dev.* 29, 197–208
25. Rieskamp, J. and Otto, P.E. (2006) SSL: a theory of how people learn to select strategies. *J. Exp. Psychol. Gen.* 135, 207–236
26. Ho, M.K. *et al.* (2022) People construct simplified mental representations to plan. *Nature* 606, 129–136
27. Lieder, F. and Griffiths, T.L. (2017) Strategy selection as rational metareasoning. *Psychol. Rev.* 124, 762–794
28. Grossmann, I. and Eibach, R.E. (2024) Metajudgment: metatheories and beliefs about good judgment across societies. *Curr. Dir. Psychol. Sci.* 33, 261–269
29. Porter, T. *et al.* (2022) Predictors and consequences of intellectual humility. *Nat. Rev. Psychol.* 1, 524–536
30. Basseches, M. (1980) Dialectical schemata: a framework for the empirical study of the development of dialectical thinking. *Hum. Dev.* 23, 400–421
31. Dong, X. *et al.* (2020) A survey on ensemble learning. *Front. Comput. Sci.* 14, 241–258
32. Horvitz, E.J. (2013) Reasoning about beliefs and actions under computational resource constraints. *arXiv* https://doi.org/10.48550/arXiv.1304.2759

33. Russell, S. and Wefald, E. (1991) Principles of metareasoning. *Artif. Intell.* 49, 361–395
34. Vilalta, R. and Drissi, Y. (2002) A perspective view and survey of meta-learning. *Artif. Intell. Rev.* 18, 77–95
35. Bilal, A. *et al.* (2025) Meta-thinking in LLMs via multi-agent reinforcement learning: a survey. *arXiv Prepr.* https://doi.org/10.48550/arXiv.2504.14520 *arXiv:2504.14520*
36. Wei, J. *et al.* (2022) Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Proces. Syst.* 35, 24824–24837
37. Yao, S. *et al.* (2023) Tree of thoughts: deliberate problem solving with large language models. *Adv. Neural Inf. Proces. Syst.* 36, 11809–11822
38. Xiang, V. *et al.* (2025) Towards system 2 reasoning in LLMs: learning how to think with meta chain-of-thought. *arXiv* https://doi.org/10.48550/arXiv.2501.04682
39. Zhou, D. *et al.* (2022) Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*
40. Wang, L. *et al.* (2023) Plan-and-solve prompting: improving zero-shot chain-of-thought reasoning by large language models. In *The 61st Annual Meeting of the Association for Computational Linguistics*
41. Wang, X. *et al.* (2023) Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*
42. Wang, Y. and Zhao, Y. (2024) Metacognitive prompting improves understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1914–1926
43. Hills, T.T. (2025) Could you be wrong: debiasing LLMs using a metacognitive prompt for improving human decision making. *arXiv* https://doi.org/10.48550/arXiv.2507.10124
44. Longo, C.F. *et al.* (2025) Eliciting metaknowledge in large language models. *Cogn. Syst. Res.* 91, 101352
45. Hsing, N. (2025) MIRROR: cognitive inner monologue between conversational turns for persistent reflection and reasoning in conversational LLMs. *arXiv* https://doi.org/10.48550/arXiv.2506.00430
46. Gaven, L. *et al.* (2022) MAGELLAN: metacognitive predictions of learning progress guide autotelic LLM agents in large goal spaces. In *Forty-second International Conference on Machine Learning*
47. Valiente, R. and Pilly, P.K. (2025) Metacognition for unknown situations and environments (MUSE). *Neural Net.*, 108131
48. Li, H. *et al.* (2025) CoRE: enhancing metacognition with label-free self-evaluation in LRMs. *arXiv* https://doi.org/10.48550/arXiv.2507.06087
49. Tan, Z. *et al.* (2025) Tuning-free accountable intervention for LLM deployment: a metacognitive approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*
50. Li, W. *et al.* (2025) Adaptive tool use in large language models with meta-cognition trigger. *arXiv* https://doi.org/10.48550/arXiv.2502.12961
51. Yan, H. *et al.* (2025) LLMs need a Bayesian meta-reasoning framework for more robust and generalizable reasoning. In *International Conference on Machine Learning*
52. Chen, H. *et al.* (2025) Pangu embedded: an efficient dual-system LLM reasoner with metacognition. *arXiv* https://doi.org/10.48550/arXiv.2505.22375
53. Ji, Y. *et al.* (2023) Metacognition-enhanced few-shot prompting with positive reinforcement. *arXiv* https://doi.org/10.48550/arXiv.2312.08642
54. Zhou, Y. *et al.* (2024) Metacognitive retrieval-augmented large language models. In *Proceedings of the ACM Web Conference*
55. Zhang, X. *et al.* (2025) MetaMind: modeling human social thoughts with metacognitive multi-agent systems. *arXiv Prepr.* https://doi.org/10.48550/arXiv.2505.18943 *arXiv: 2505.18943*
56. Lee, J. *et al.* (2025) Pragmatic metacognitive prompting improves LLM performance on sarcasm detection. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pp. 63–70
57. Toy, J. *et al.* (2024) Metacognition is all you need? Using introspection in generative agents to improve goal-directed behavior. *arXiv* https://doi.org/10.48550/arXiv.2401.10910
58. Steyvers, M. and Peters, M.A. (2025) Metacognition and uncertainty communication in humans and large language models. *arXiv* https://doi.org/10.48550/arXiv.2504.14045
59. Ji-An, L. *et al.* (2025) Language models are capable of metacognitive monitoring and control of their internal activations. *arXiv* https://doi.org/10.48550/arXiv.2505.13763
60. Didolkar, A. *et al.* (2024) Metacognitive capabilities of LLMs: an exploration in mathematical problem solving. *Adv. Neural Inf. Proces. Syst.* 37, 19783–19812
61. Ma, Z. *et al.* (2025) Large language models have intrinsic metacognition, but need a good lens. *arXiv* https://doi.org/10.48550/arXiv.2506.08410
62. Pavlovic, J. *et al.* (2024) Generative AI as a metacognitive agent: a comparative mixed-method study with human participants on ICF-mimicking exam performance. *arXiv* https://doi.org/10.48550/arXiv.2405.05285
63. Minaee, S. *et al.* (2024) Large language models: a survey. *arXiv* https://doi.org/10.48550/arXiv.2402.06196
64. Li, Y. *et al.* (2024) I think, therefore I am: benchmarking awareness in large language models using AwareBench. *arXiv* https://doi.org/10.48550/arXiv.2401.17882
65. Cash, T.N. *et al.* (2025) Quantifying uncert-AI-nty: testing the accuracy of LLMs' confidence judgments. *Mem. Cogn.* https://doi.org/10.3758/s13421-025-01755-4
66. Scholten, F. *et al.* (2024) Metacognitive myopia in large language models. *arXiv* https://doi.org/10.48550/arXiv.2408.05568
67. Dubova, M. *et al.* (2025) Is Ockham's razor losing its edge? New perspectives on the principle of model parsimony. *Proc. Natl. Acad. Sci. U. S. A.* 122, e2401230121
68. Carruthers, P. (2009) Mindreading underlies metacognition. *Behav. Brain Sci.* 32, 164–182
69. Chater, N. (2018) *The Mind Is Flat: The Remarkable Shallowness of the Improvising Brain*, Yale University Press
70. Cushman, F. (2020) Rationalization is rational. *Behav. Brain Sci.* 43, e28
71. Nisbett, R.E. and Wilson, T.D. (1977) Telling more than we can know: verbal reports on mental processes. *Psychol. Rev.* 84, 231–259
72. Chen, Y. *et al.* (2025) Reasoning models don't always say what they think. *Anthropic Res.* https://doi.org/10.48550/arXiv.2505.05410
73. Ohtani, R. *et al.* (2024) Does metacognitive prompting improve causal inference in large language models? In *IEEE Conference on Artificial Intelligence*
74. Collins, K.M. *et al.* (2024) Building machines that learn and think with people. *Nat. Hum. Behav.* 8, 1851–1863
75. Gopnik, A. and Wellman, H.M. (1992) Why the child's theory of mind really is a theory. *Mind Lang.* 7, 145–171
76. Chater, N. *et al.* (2018) Negotiating the traffic: can cognitive science help make autonomous vehicles a reality? *Trends Cogn. Sci.* 22, 93–95
77. Gallese, V. and Goldman, A. (1998) Mirror neurons and the simulation theory of mind-reading. *Trends Cogn. Sci.* 2, 493–501
78. Baker, C.L. *et al.* (2009) Action understanding as inverse planning. *Cognition* 113, 329–349
79. Sperber, D. *et al.* (2010) Epistemic vigilance. *Mind Lang.* 25, 359–393
80. Sobel, D.M. and Kushnir, T. (2013) Knowledge matters: how children evaluate the reliability of testimony as a process of rational inference. *Psychol. Rev.* 120, 779–797
81. Mercier, H. and Sperber, D. (2017) *The Enigma of Reason*, Harvard University Press
82. Fehr, E. and Fischbacher, U. (2004) Third-party punishment and social norms. *Evol. Hum. Behav.* 25, 63–87
83. Frank, R.H. (1988) *Passions within Reason: The Strategic Role of the Emotions*, Norton
84. Grossmann, I. *et al.* (2017) Wise deliberation sustains cooperation. *Nat. Hum. Behav.* 1, 0061
85. Peetz, J. and Grossmann, I. (2021) Wise reasoning about the future is associated with adaptive interpersonal feelings after relational challenges. *Soc. Psychol. Personal. Sci.* 12, 629–637

86. Dalrymple, D. *et al.* (2024) Towards guaranteed safe AI: a framework for ensuring robust and reliable AI systems. *arXiv* https://doi.org/10.48550/arXiv.2405.06624

87. Johnson, B. (2022) Metacognition for artificial intelligence system safety: an approach to safe and desired behavior. *Saf. Sci.* 151, 105743

88. Bostrom, N. (2014) *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press

89. Melkonyan, T. *et al.* (2022) The cognitive foundations of tacit commitments: a virtual bargaining model of dynamic interactions. *J. Math. Psychol.* 108, 102640

90. Sagiv, L. and Schwartz, S.H. (2022) Personal values across cultures. *Annu. Rev. Psychol.* 73, 517–546

91. Varnum, M.E. and Grossmann, I. (2017) Cultural change: the how and the why. *Perspect. Psychol. Sci.* 12, 956–972

92. Rawls, J. (1971) *A Theory of Justice*, Harvard University Press

93. Heyes, C. *et al.* (2020) Knowing ourselves together: the cultural origins of metacognition. *Trends Cogn. Sci.* 24, 349–362

94. Hu, C.S. *et al.* (2023) Resting state default mode network is associated with wise advising. *Sci. Rep.* 13, 14239

95. Hutchinson, J.M. and Gigerenzer, G. (2005) Simple heuristics and rules of thumb: where psychologists and behavioural biologists might meet. *Behav. Process.* 69, 97–124

96. Dong, M. *et al.* (2023) Thirty years of psychological wisdom research: what we know about the correlates of an ancient concept. *Perspect. Psychol. Sci.* 18, 778–811

97. Henrich, J. (2018) *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*, Princeton University Press

98. Lampinen, A.K. *et al.* (2022) Tell me why! explanations support learning relational and causal structure. In *International Conference on Machine Learning*, pp. 11868–11890

99. Lindsey, J. *et al.* (2025) On the biology of a large language model. *Anthropic Res.* https://doi.org/10.59350/q8msb-1s677

100. Ritter, F.E. *et al.* (2019) ACT-R: a cognitive architecture for modeling cognition. *Wiley Interdiscip. Rev. Cogn. Sci.* 10, e1488

101. Laird, J.E. *et al.* (1987) Soar: an architecture for general intelligence. *Artif. Intell.* 33, 1–64

102. Goertzel, B. *et al.* (2023) OpenCog Hyperon: a framework for AGI at the human level and beyond. *arXiv* https://doi.org/10.48550/arXiv.2310.18318

103. Vezhnevets, A.S. *et al.* (2017) FeUdal networks for hierarchical reinforcement learning. In *International Conference on Machine Learning*, pp. 3540–3549

104. Zhang, K. *et al.* (2021) Multi-agent reinforcement learning: a selective overview of theories and algorithms. In *Handbook of Reinforcement Learning and Control* (Vamvoudakis, K.G. *et al.*, eds), pp. 321–384

105. Hassan, S. and De Filippi, P. (2021) Decentralized autonomous organization. *Internet Policy Rev.* 10

106. Irving, G. *et al.* (2018) AI safety via debate. *arXiv* https://doi.org/10.48550/arXiv.1805.00899

107. Bai, Y. *et al.* (2022) Constitutional AI: harmlessness from AI feedback. *arXiv* https://doi.org/10.48550/arXiv.2212.08073

108. Huang, C. *et al.* (2020) Stronger resting-state neural oscillations associated with wiser advising from the 2nd-but not the 3rd-person perspective. *Sci. Rep.* 10, 12677

109. Wang, G. *et al.* (2025) Decoupling metacognition from cognition: a framework for quantifying metacognitive ability in LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*

110. Fleming, S.M. (2023) Metacognitive psychophysics in humans, animals, and AI: a research agenda for mapping introspective systems. *J. Conscious. Stud.* 30, 113–128

111. Fleming, S.M. and Daw, N.D. (2017) Self-evaluation of decision-making: a general Bayesian framework for metacognitive computation. *Psychol. Rev.* 124, 91–114

112. Fleming, S.M. and Lau, H.C. (2014) How to measure metacognition. *Front. Hum. Neurosci.* 8, 443

113. Strachan, J.W. *et al.* (2024) Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* 8, 1285–1295

114. Webb, T. *et al.* (2023) Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* 7, 1526–1541

115. Frank, M.C. (2023) Baby steps in evaluating the capacities of large language models. *Nat. Rev. Psychol.* 2, 451–452

116. Gandhi, K. *et al.* (2023) Understanding social reasoning in language models with language models. *Adv. Neural Inf. Proces. Syst.* 36, 13518–13529

117. Thagard, P. (2024) Can ChatGPT make explanatory inferences? Benchmarks for abductive reasoning. *arXiv* https://doi.org/10.48550/arXiv.2404.18982

118. Grossmann, I. *et al.* (2010) Reasoning about social conflicts improves into old age. *Proc. Natl. Acad. Sci. U. S. A.* 107, 7246–7250

119. Stavropoulos, A. *et al.* (2024) Shadows of wisdom: classifying meta-cognitive and morally grounded narrative content via large language models. *Behav. Res. Methods* 56, 7632–7646

120. Lieder, F. and Griffiths, T.L. (2020) Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* 43, e1

121. Simon, H.A. (1955) A behavioral model of rational choice. *Q. J. Econ.* 69, 99–118

122. Levine, S. *et al.* (2024) Resource-rational contractualism: a triple theory of moral cognition. *Behav. Brain Sci.* Published online September 26, 2005. https://doi.org/10.1017/s0140525x24001067

123. Sanborn, A.N. and Chater, N. (2016) Bayesian brains without probabilities. *Trends Cogn. Sci.* 20, 883–893

124. Hayek, F.A. (1945) The use of knowledge in society. *Am. Econ. Rev.* 35, 519–530

125. Kitcher, P. (1990) The division of cognitive labor. *J. Philos.* 87, 5–22

126. Burton, J.W. *et al.* (2024) How large language models can reshape collective intelligence. *Nat. Hum. Behav.* 8, 1643–1655