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], Igor Grossmann[1*]

1 University of Waterloo, Department of Psychology
2 University of Waterloo, Department of Electrical and Computer Engineering
3 Université de Montréal, Department of Computer Science and Operations Research
4 Warwick Business School, Behavioural Science Group
5 Stanford University, Department of Psychology
6 University of Waterloo, Cheriton School of Computer Science
7 Google DeepMind
8 Santa Fe Institute
9 Max Planck Institute for Human Development
10 Max Planck Institute for Intelligent Systems

*** Correspondence to**: Sam Johnson (samuel.johnson@uwaterloo.ca) *or* Igor Grossmann (igrossma@uwaterloo.ca)

Word count: 3742 (main text)

Last edited: May 6, 2025 (submitted to journal)

Note. This article is a preprint. It has not been peer-reviewed and, upon publication, will be superseded by the accepted journal article as the version of record.

Highlights

- We examine the why and the how of building wise AI
- Wisdom helps humans to navigate intractable problems through object-level strategies (for managing problems) and metacognitive strategies (for managing object-level strategies)
- Wise AI, through improved metacognition, 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 AI architecture for metacognition

Abstract

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

Keywords: AI, wisdom, metacognition, reasoning, decision-making

Imagining and building wise machines: The centrality of AI metacognition

Ongoing challenges

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

Here, we argue that AIs lack a key capability that underlies all these deficiencies: they are not **wise**.

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 to them 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 much more junior colleague when she recognizes that the colleague 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; **Table 1**]. In our view, differences across theories mask important generalizations about wisdom's function and mechanisms (see [4,11] for more detail).

Theory	Elements of Wisdom
	Component Theories
Balance Theory [10]	Deploying knowledge and skills to achieve the common good by: - **Balancing interests** (their own, others', and society's) - **Balancing time perspectives** (long-term and short-term) - **Deploying positive ethical values** - **Managing environments** (adapting to, selecting, or altering)
Berlin Wisdom Model [6]	Expertise in important and difficult matters of life: - **Factual knowledge** (about human nature and life) - **Procedural knowledge** (strategies to address life challenges) - **Contextualism** (strategies account for social context) - **Value relativism** (strategies account for variation in values) - **Managing uncertainty** (strategies change with circumstances)
MORE Life Experience Model [7]	Gaining psychological resources via reflection, to cope with life challenges: - **Uncertainty management** (coping with uncertainty, uncontrollability) - **Openness** (to new experiences and perspectives) - **Reflectivity** (about life experiences) - **Emotion regulation** (management of and sensitivity to emotions)
Three-Dimensional Model [5]	Acquiring and reflecting on life experience to cultivate personality traits: - **Cognitive** (curiosity about life; recognizing uncertainty, ignorance) - **Emotional** (sympathy and compassion; valuing others) - **Reflective** (perspective-taking; questioning one's beliefs)
Wise Reasoning Model [9]	Using context-sensitive reasoning to manage important social challenges: - **Intellectual humility** (knowledge of one's epistemic limits) - **Perspective-taking** (actively seeking out others' viewpoints) - **Perspective integration** (accounting for multiple perspectives) - **Flexibility** (recognizing uncertainty and change)
	Consensus Models
Common Wisdom Model [4]	A style of social-cognitive processing that is: - **Morally grounded** o Balancing interests of the self and others o Pursuing truth o Oriented toward the common good - **Metacognitively sound** o Considering context o Taking multiple perspectives o Accounting for short- and long-term effects o Thinking reflectively o Aware of the limits of one's knowledge
Integrative Model [8]	A behavioral repertoire in which: - A complex and uncertain **situation** arises, evoking an appropriate **emotional** and **motivational state** o Open-mindedness, care for others, calm emotions - Depending on **traits and skills** o Exploratory orientation, concern for others, emotion regulation - Facilitating deployment of **cognitive resources** o Life knowledge, metacognition, reflection - Using these resources to deploy effective **metacognitive strategies** o Reasoning is contextualized, balanced, multi-perspectival

Table 1. Psychological approaches to wisdom. The five "component theories" are a selected set of psychological theories of wisdom. The two "consensus models" are attempts to identify common themes and processes among those theories. For a more detailed review, see [8].

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 those answers. Natural selection would have made us nothing more or less than master statisticians, merciless optimizers, lightning calculators. Indeed, in some domains—like low-level visual processing—we approximate this ideal.

Yet, social interaction and decision-making in an unstructured and ever-changing world require further tools [12]. Such problems are often **intractable** in one or more ways:

- *Incommensurable*. Conflicting values are at stake that cannot be put on the same scale [13].
- *Transformative*. The outcome of the decision changes one's preferences, creating a clash between present and future values [14].
- *Radically uncertain*. One cannot exhaustively list possible outcomes or non-arbitrarily assign probabilities [15].
- *Chaotic*. The data-generating process has a strong nonlinearity or dependency on initial conditions, making it fundamentally unpredictable [16].
- *Non-stationary*. The underlying process changes over time, making the probability distribution unlearnable.
- *Out-of-distribution*. The situation is far beyond one's experience or available data.
- *Computationally explosive*. The optimal response could be calculated only with infeasibly large computational resources.

Our earlier examples of wisdom featured such intractability. Wisdom helped Willa understand how to make an incommensurable trade-off, Daphne to navigate her ignorance in an out-of-distribution situation, and Ron to 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 (**Figure 1**): **Object-level strategies** to manage the problem itself (i.e., the "object" of judgment) and **metacognitive strategies** to manage those object-level strategies, particularly when they conflict [17-18]. We sketch this view here, providing a more detailed defense elsewhere [4,11].

Object-level strategies often take the form of **heuristics**—rules of thumb which rely on a small number of inputs and do not attempt to execute a complex analysis [19] but may approximate it [20]. For example, Willa may have used a heuristic like "Prioritize family relationships" to help her children, and Ron may have used a heuristic like "Avoid the worst-case scenario" to help his candidate. Heuristics often work well, despite requiring less computation than optimization, because they focus on just the most relevant information, reducing the chances of overfitting [19]. Much of "folk wisdom" comprises culturally-evolved heuristics, transmitted across generations (e.g., the heuristic to defer to elders).

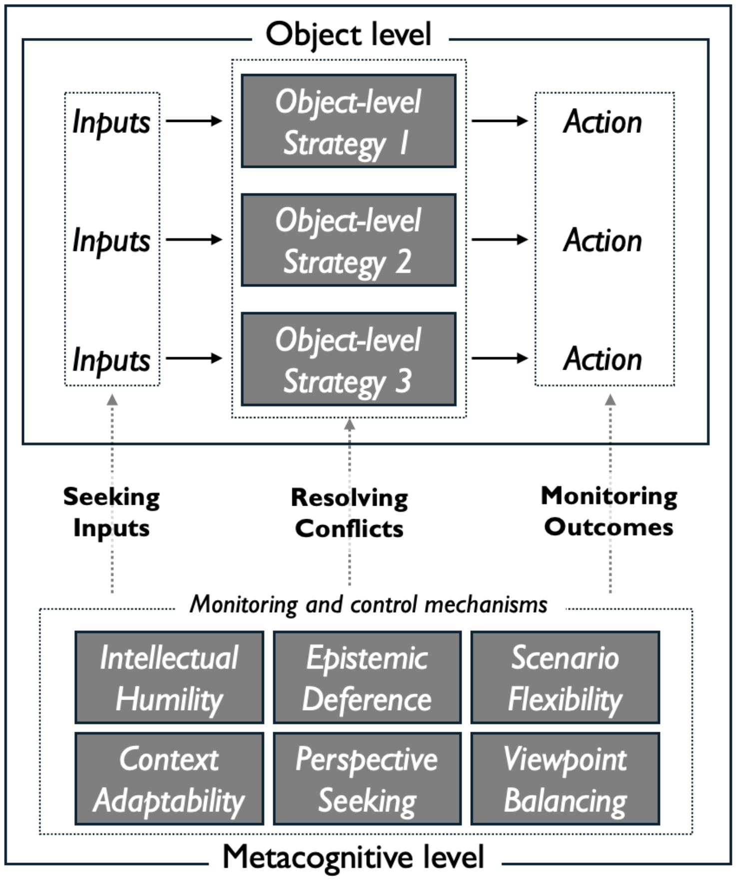

Figure 1. The relationship between object-level and metacognitive strategies in wise reasoning. Object-level strategies (e.g., heuristics, narratives, analytical procedures) provide candidate actions for a given situation. Metacognitive monitoring and control processes regulate these strategies in three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

Another type of object-level strategy is **narrative** thinking—using causal knowledge and analogies to construct a mental model that can explain a situation, generate predictions, and evaluate choices [12,21]. When Ron constructs worst-case scenarios, he uses his causal knowledge (about government policy and voter psychology) and comparable experiences (about the fates of other campaigns). Like heuristics, narratives can be socially transmitted and adapted within and across generations [22-23] (e.g., the Protestant Work Ethic narrative).

Sound object-level strategies, however, are insufficient for wisdom:

- 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 to fill any gaps.)

- Strategies often yield conflicting advice; a **conflict resolution process** is required to select the best strategy for each situation [24]. (Should Daphne follow the strategy “trust your judgment” or the equally-plausible “trust knowledgeable experts”?)
- Strategies can break under unfavorable conditions, as when the underlying pattern changes unpredictably; an **outcome-monitoring process** is required to safeguard 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 [25-27]—using metacognitive strategies (**Table 2**) [4]. Daphne exhibits intellectual humility when she recognizes that she does not understand her patient’s symptoms; perspective-seeking when she calls upon her colleague’s expertise; context adaptability when she considers whether her patient’s unique situation limits the relevance of her colleague’s expertise; and ultimately epistemic deference when she adopts her colleague’s view.

Metacognitive Process	Description
Intellectual humility	Awareness of what one does and does not know; acknowledgment of uncertainty and one’s fallibility [83]
Epistemic deference	Willingness to defer to others’ expertise when appropriate [84]
Scenario flexibility	Considering diverse ways in which a scenario might unfold to identify possible contingencies
Context adaptability	Identifying features of a situation that make it comparable to or distinct from other situations [6]
Perspective seeking	Drawing on multiple perspectives where each offers information for reaching a good decision [6]
Viewpoint balancing	Recognizing and integrating discrepant interests [10,71]

Table 2. Example metacognitive processes commonly exhibited by wise people. For more detail, see [4] and [8] (especially Table 1).

Toward wise AI: Machine metacognition

Could metacognitive AI—with the ability to model its own computations and use that model to optimize subsequent computations—help machines to perform better in intractable situations?

Although AI metacognition has precedents [28-31], existing research has focused on object-level strategies like heuristics [32]. GenAI models can perform well in some metacognitive tasks (e.g., classifying math problems by solving procedure [33]) and state-of-the-art models exhibit rudimentary forms of metacognition (e.g., using an inference-time search to decide when to stop searching). Yet they fail at more complex metacognition. They often “hallucinate” an answer rather than admit ignorance [34] and they struggle to understand their goals [35], capabilities [35], and strength of their evidence [36]—symptoms of a broader “metacognitive myopia” [37].

Would a wise AI think like a wise human? Perhaps not. Much of human metacognition is adapted for economizing scarce cognitive resources [19,38-39], and many biases may be side-effects of solving this constrained optimization problem [40-41]. 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 our cognition to the social environment (as in the division of physical or cognitive labor [42-43]), 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 to an extensive, integrated knowledge base.

Conversely, perhaps AI wisdom would converge considerably with human wisdom. AI wisdom also faces computational constraints, since compute can be costly. Moreover, heuristics work for AI for the same reasons 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 [44].

What are the potential benefits of wise AI?

Robustness

Given the range of intractable environments in which intelligent systems must operate, three failures of robustness are common:

- *Unreliability.* Given similar inputs, a system can produce wildly different outputs. This could be caused by applying different strategies each time, or applying a strategy that produces inconsistent results.
- *Bias*. The output is systematically wrong or non-representative in a predictable direction.
- *Inflexibility.* Novel inputs lead to lower-quality outputs.

Human wisdom combines object-level and metacognitive strategies to adapt robustly across environments. Object-level strategies like heuristics are beneficial because they outperform analytic optimization by avoiding data overfitting [19,45], especially in novel, out-of-distribution contexts. These strategies are supported by wise metacognition, which helps reasoners to learn new information from other perspectives and discern its relevance, to balance the competing urges to simplify and optimize, and to avoid catastrophic error by checking the plausibility of a strategy's output.

For similar reasons, wise AI would be more robust 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, identifying sample deficiencies in its training data (perhaps requesting additional data), and understanding the causal process by which biases resulted (correcting for that bias). 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. Even worse, AI can confabulate false explanations for their outputs [46]. Explainability is thus a focus in AI research [1]. Although cognitive scientists disagree about the extent of introspective access in humans [47], 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. Explainability comes "for free" with metacognition.

Alternatively, the mind may be "flat" [48]—it does not contain hidden depths of reasons that can be uncovered through introspection. When we report our thought processes, we are reporting *inferences* ("stories"), not observations. The reasoner observes the outputs of her strategies and reasons backwards to what could have caused them [49]. These inferences may often be incorrect [50], yet they are often useful and, when verbally formulated, 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.

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), and influencing human–human cooperation (e.g., social media content curation). Cooperative AI [2,51] 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 [52], including the tacit ability to form joint plans to coordinate behavior [53]. In humans, this is accomplished through object-level strategies such as first-person simulation (putting oneself in the other's shoes) [54] and third-person, theory-based reasoning (e.g., assuming that the agent is rational [55]).

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 [56]. Even young children develop object-level strategies for evaluating sources—tracking cues such as accurate past testimony and conflicts of interest [57]—and more sophisticated reasoners can check whether the reasoning itself is valid [58]. Such "epistemic vigilance" mechanisms make credible communication among humans possible: Without a means of assessing a 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 [59], while emotions like shame and guilt impose internal costs [60]. Humans sharing a cultural and psychological context can assume these costs as common ground, promoting credible commitment.

Wise metacognition is required to effectively manage these object-level mechanisms [61-62]—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 [3,63]. For now, the main safety risks are simply that systems that we come to rely on fail us—a shoddy surgical robot, incompetent tax advice, or biased parole algorithm. Machine metacognition can help to avoid such failures [64]. AIs with well-calibrated confidence can target the most likely risks; appropriate self-models would help AIs to anticipate failures; and continual monitoring of its 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 [65]. 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 [65] illustrates both points in his parable of the paperclip-maximizing AI who converts the Earth into paperclips 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 unspoken assumptions we make and which an AI would not necessarily share. 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 [66].

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 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., cross-cultural and religious differences in maximizing happiness vs. liberty) [67]. 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 [68]? Aligning AI to current values would risk reifying those values as “the right” values, stalling future social progress.

A two-pronged, wisdom-oriented approach may be more promising.

First, AIs must themselves implement wise reasoning—aligning them to the right object-level and metacognitive strategies rather than to 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 like 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, our 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 [69]—bringing situation-specific judgments and general moral principles into alignment with one another through iterative adjustments.

How might we build wise AI?

Benchmarking

Wisdom is difficult to **benchmark**. Wisdom is context-sensitive, so the benchmark input must contain sufficient detail to match the rich context of a real-world situation. Moreover, since wisdom is about the reasoning underlying strategy selection, the benchmark must evaluate not only the outcome but the process that led to it.

To make progress, let's consider how other complex constructs have been benchmarked. One approach is to collect tasks from psychology experiments, akin to benchmarking theory-of-mind or analogical reasoning [70-71]. 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 [72-73]. 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 [74]—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 benchmark AI wisdom would start with tasks that measure wise reasoning in humans [75]. These tasks present participants with a social dilemma or a choice between seemingly incommensurable options, asks them to reflect on the next steps, with reflections scored on prespecified criteria by human raters. 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) [76]. 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 about themselves.

Ultimately, the wisdom of increasingly autonomous AIs, as with people, will be judged by the rest of us. 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.

Training

Training object-level and metacognitive wisdom may require different strategies.

In humans, object-level strategies like 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 [77] and from others [78] may be more promising.

Yet this approach is unlikely to work for training metacognition, where the challenge is deciding between strategies in a context-sensitive way. This contrasts with typical AI training, where a loss function defined over the model's outputs (rather than reasoning) is minimized. 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.

This problem may require multiple complementary approaches. One possibility is a two-step process: first training models for wise strategy selection directly (e.g., correctly identifying when to be intellectually humble) and then training them to use those strategies correctly (e.g., carrying out intellectual humble behavior). A second possibility is to evaluate whether models can plausibly explain their metacognitive strategies in benchmark cases, and then simultaneously train strategies and outputs (e.g., training the model to identify the situation as one that calls for intellectual humility *and* to reason accordingly [79]). In either case, models could be trained against what a wise human would do or against the acceptability of its explanations for its choices.

Architecture

LLMs work by generating the next token (i.e., word or word part) based on the input in its **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. The great discovery has been that this process can yield surprisingly intelligent outputs—and even some degree of planning (e.g., planning rhymes in a poem [80])—given a large enough network and enough training data. Yet, given their lack of explicit planning, perhaps it is unsurprising that LLMs struggle with metacognition, 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. **Table 3** lists some possible ways to engineer metacognition, some of which have precedents. For example, in "chain-of-thought" prompting, the model produces intermediate reasoning steps, which often leads to improved performance [81]. The more recent "meta chain-of-thought" framework [82] suggests how this technique can be extended to improve reasoning for difficult problems that require backtracking and branching, in turn demanding greater metacognitive control.

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 (entropy, calibration error) and coherence metrics (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 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 [85], SOAR [86]), where a metacognitive layer explicitly monitors and selects object-level strategies. Develop explicit reflective subsystems 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: a) Hybrid symbolic/sub-symbolic approaches (e.g., OpenCog Hyperon [87], ACT-R style modules); b) Reinforcement learning hierarchical controllers (e.g., FeUdal networks [88]) Introduce standalone "auditor" modules trained explicitly to critique primary outputs for internal consistency, logical coherence, or sensitivity to constraints. For instance, chain-of-thought prompting with GPT-4 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 [81] approaches, which generate explicit narration of metacognitive reasoning steps in understandable language. *Interactive debugging & 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 multi-agent reasoning and collective decision-making, analogous to human reliance on socially distributed cognition. Implement epistemic cross-checking and adversarial debate between multiple AI systems to mitigate individual AI overconfidence and misinformation propagation.	*Multi-agent epistemic vigilance:* Multiple independent AI agents work collaboratively, requiring agreement or consensus for outputs on critical tasks. *Concrete architectures:* Multi-agent RL (MARL) [89], decentralized autonomous organizations (DAO)-inspired decision-making [90]. *Debate-based metacognitive cross-checking:* AI reasoning outputs must pass adversarial debates or cross-examinations from independently trained AI debaters before being finalized. *Example frameworks*: OpenAI's debate-style AI safety approach [91], Anthropic's Constitutional AI approach [92].

Table 3. Engineering wiser AI via metacognition.

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 [61]. Perhaps wise AI would have these qualities, too.

Yet 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 two responses.

First, if the alternative were halting all AI progress, building wise AI would introduce added risks. But compared to 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, 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 (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.

Outstanding Questions

- How might wise 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?
- 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 offloading of metacognitive labor, leading to 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?
- 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 of AI itself?
- What further considerations would be required to embody metacognition in robots?

References

[1] Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., ... & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*, *55*, 1-33.

[2] Dafoe, A., Hughes, E., Bachrach, Y., Collins, T., McKee, K. R., Leibo, J. Z., Larson, K., & Graepel, T. (2020). Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*.

[3] Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., ... & Gao, W. (2023). AI alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

[4] Grossmann, I., Weststrate, N. M., Ardelt, M., Brienza, J. P., Dong, M., Ferrari, M., ... & Vervaeke, J. (2020). The science of wisdom in a polarized world: Knowns and unknowns. *Psychological Inquiry*, *31*, 103–133.

[5] Ardelt, M. (2004). Wisdom as expert knowledge system: a critical review of a contemporary operationalization of an ancient concept. *Human Development*, *47*, 257–287.

[6] Baltes, P. B., & Smith, J. (2008). The fascination of wisdom: Its nature, ontogeny, and function. *Perspectives on Psychological Science*, *3*, 56–64.

[7] Glück, J., & Bluck, S. (2013). The MORE Life Experience Model: A theory of the development of personal wisdom. In M. Ferrari & N. M. Weststrate (Eds.), *The scientific study of personal wisdom* (pp. 75–98). Berlin, Germany: Springer.

[8] Glück, J., & Weststrate, N. M. (2022). The wisdom researchers and the elephant: An integrative model of wise behavior. *Personality and Social Psychology Review*, *26*, 342–374.

[9] Grossmann, I. (2017). Wisdom in context. *Perspectives on Psychological Science*, *12*, 233–257.

[10] Sternberg, R. J. (1998). A balance theory of wisdom. *Review of General Psychology*, *2*, 347–365.

[11] Grossmann, I., & Johnson, S. (2025). Cultivating wisdom through metacognition: A new frontier in decision-making under radical uncertainty. *PsyArXiv preprint hcnkx_v2.*

[12] Johnson, S. G. B., Bilovich, A., & Tuckett, D. (2023). Conviction narrative theory: A theory of choice under radical uncertainty. *Behavioral and Brain Sciences*, *46*, e82.

[13] Walasek, L., & Brown, G. D. (2023). Incomparability and incommensurability in choice: No common currency of value? *Perspectives on Psychological Science*, 17456916231192828.

[14] Paul, L. A. (2013). *Transformative experience*. Oxford, UK: Oxford University Press.

[15] Kay, J., & King, M. (2020). *Radical uncertainty: Decision-making beyond the numbers*. New York, NY: Norton.

[16] Lorenz, E. (1993). *The essence of chaos*. Seattle, WA: University of Washington Press.

[17] Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation* (Vol. 26, pp. 125–173). Cambridge, MA: Academic Press.

[18] Karlan, B., & Allen, C. (2024). Engineered wisdom for learning machines. *Journal of Experimental & Theoretical Artificial Intelligence*, *36*, 257-272.

[19] Todd, P. M., & Gigerenzer, G. (2012). *Ecological rationality: Intelligence in the world*. New York, NY: Oxford University Press.

[20] Parpart, P., Jones, M., & Love, B. C. (2018). Heuristics as Bayesian inference under extreme priors. *Cognitive Psychology*, *102*, 127-144.

[21] Glück, J., Bluck, S., Baron, J., & McAdams, D. (2005). The wisdom of experience: Autobiographical narratives across adulthood. *International Journal of Behavioral Development*, *29*, 197–208.

[22] Edmondson, R., & Woerner, M. H. (2019). Sociocultural foundations of wisdom. In R. J. Sternberg & J. Glück (Eds.), *The Cambridge handbook of wisdom* (pp. 40-68). Cambridge, UK: Cambridge University Press.

[23] Gigerenzer, G. (2023). How do narratives relate to heuristics? *Behavioral and Brain Sciences*, *46*, e94.

[24] Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135*, 207–236.

[25] Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2022). People construct simplified mental representations to plan. *Nature*, *606*, 129-136.

[26] Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review, 124*, 762–794.

[27] Grossmann, I., & Eibach, R. E. (2024). Metajudgment: Metatheories and beliefs about good judgment across societies. *Current Directions in Psychological Science.*

[28] Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*, 241–258.

[29] Horvitz, E. J. (2013). Reasoning about beliefs and actions under computational resource constraints. *arXiv preprint arXiv:1304.2759*.

[30] Russell, S., & Wefald, E. (1991). Principles of metareasoning. *Artificial Intelligence*, *49*, 361–395.

[31] Vilalta, R., & Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial Intelligence Review*, *18*, 77-95.

[32] Pearl, J. (1984). *Heuristics: Intelligent search strategies for computer problem solving*. Boston, MA: Addison-Wesley.

[33] Didolkar, A., Goyal, A., Ke, N. R., Guo, S., Valko, M., Lillicrap, T., ... & Arora, S. (2024). Metacognitive capabilities of LLMs: An exploration in mathematical problem solving. *arXiv preprint arXiv:2405.12205*.

[34] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

[35] Li, Y., Huang, Y., Lin, Y., Wu, S., Wan, Y., & Sun, L. (2024). I think, therefore I am: Awareness in Large Language Models. *arXiv preprint arXiv:2401.17882*.

[36] Cash, T. N., Oppenheimer, D. M., & Christie, S. (2024). Quantifying UncertAInty: Testing the Accuracy of LLMs' Confidence Judgments. *PsyArXiv preprint 47df5_v1*.

[37] Scholten, F., Rebholz, T. R., & Hütter, M. (2024). Metacognitive myopia in Large Language Models. *arXiv preprint arXiv:2408.05568*.

[38] Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1.

[39] Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*, 99–118.

[40] Levine, S., Chater, N., Tenenbaum, J., & Cushman, F. (2024). Resource-rational contractualism: A triple theory of moral cognition. *Behavioral and Brain Sciences*.

[41] Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*, 883-893.

[42] Hayek, F. A. (1945). The use of knowledge in society. *American Economic Review*, *35*, 519–530.

[43] Kitcher, P. (1990). The division of cognitive labor. *The Journal of Philosophy*, *87*, 5-22.

[44] Burton, J. W., Lopez-Lopez, E., Hechtlinger, S., Rahwan, Z., Aeschbach, S., Bakker, M. A., ... & Hertwig, R. (2024). How large language models can reshape collective intelligence. *Nature Human Behaviour*, *8*, 1643-1655.

[45] Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, *81*, 95–106.

[46] Chen, Y., Benton, J., Radhakrishnan, A., Denison, J. U. C., Schulman, J., Somani, A., ... & Perez, E. (2025). Reasoning models don't always say what they think. *Anthropic Research*.

[47] Carruthers, P. (2009). Mindreading underlies metacognition. *Behavioral and Brain Sciences*, *32*, 164–182.

[48] Chater, N. (2018). *The mind is flat: The remarkable shallowness of the improvising brain*. New Haven, CT: Yale University Press.

[49] Cushman F. (2020) Rationalization is rational. *Behavioral and Brain Sciences*, *43*, e28.

[50] Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259.

[51] Collins, K. M., Sucholutsky, I., Bhatt, U., Chandra, K., Wong, L., Lee, M., ... & Griffiths, T. L. (2024). Building machines that learn and think with people. *Nature Human Behaviour*, *8*, 1851–1863.

[52] Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really *is* a theory. *Mind & Language*, *7*, 145–171.

[53] Chater, N., Misyak, J., Watson, D., Griffiths, N., & Mouzakitis, A. (2018). Negotiating the traffic: Can cognitive science help make autonomous vehicles a reality? *Trends in Cognitive Sciences*, *22*, 93–95.

[54] Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*, 493-501.

[55] Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*, 329–349.

[56] Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*, 359-393.

[57] Sobel, D. M., & Kushnir, T. (2013). Knowledge matters: How children evaluate the reliability of testimony as a process of rational inference. *Psychological Review, 120*, 779–797.

[58] Mercier, H., & Sperber, D. (2017). *The enigma of reason*. Cambridge, UK: Harvard University Press.

[59] Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*, 63-87.

[60] Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions*. New York, NY: Norton.

[61] Grossmann, I., Brienza, J. P., & Bobocel, D. R. (2017). Wise deliberation sustains cooperation. *Nature Human Behaviour*, *1*, 0061.

[62] Peetz, J., & Grossmann, I. (2021). Wise reasoning about the future is associated with adaptive interpersonal feelings after relational challenges. *Social Psychological and Personality Science*, *12*, 629-637.

[63] Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., ... & Tenenbaum, J. (2024). Towards guaranteed safe AI: A framework for ensuring robust and reliable AI systems. arXiv preprint arXiv:2405.06624.

[64] Johnson, B. (2022). Metacognition for artificial intelligence system safety: An approach to safe and desired behavior. *Safety Science*, *151*, 105743.

[65] Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford, UK: Oxford University Press.

[66] Melkonyan, T., Zeitoun, H., & Chater, N. (2022). The cognitive foundations of tacit commitments: A virtual bargaining model of dynamic interactions. *Journal of Mathematical Psychology*, *108*, 102640.

[67] Sagiv, L., & Schwartz, S. H. (2022). Personal values across cultures. *Annual Review of Psychology*, *73*, 517–546.

[68] Varnum, M. E., & Grossmann, I. (2017). Cultural change: The how and the why. *Perspectives on Psychological Science*, *12*, 956-972.

[69] Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.

[70] Strachan, J. W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., ... & Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *8*, 1285–1295.

[71] Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*, 1526–1541.

[72] Frank, M. C. (2023). Baby steps in evaluating the capacities of large language models. *Nature Reviews Psychology*, *2*, 451–452.

[73] Gandhi, K., Fränken, J. P., Gerstenberg, T., & Goodman, N. (2023). Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, *36*, 13518–13529.

[74] Thagard, P. (2024). Can ChatGPT make explanatory inferences? Benchmarks for abductive reasoning. *arXiv preprint arXiv:2404.18982*.

[75] Grossmann, I., Na, J., Varnum, M. E. W., Park, D. C., Kitayama, S., & Nisbett, R. E. (2010). Reasoning about social conflicts improves into old age. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 7246–7250.

[76] Stavropoulos, A., Crone, D. L., & Grossmann, I. (2024). Shadows of wisdom: Classifying meta-cognitive and morally grounded narrative content via large language models. *Behavior Research Methods*, *56*, 7632–7646.

[77] Dong, M., Weststrate, N. M., & Fournier, M. A. (2023). Thirty years of psychological wisdom research: What we know about the correlates of an ancient concept. *Perspectives on Psychological Science*, *18*, 778–811.

[78] Henrich, J. (2018). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton, NJ: Princeton University Press.

[79] Lampinen, A. K., Roy, N., Dasgupta, I., Chan, S. C., Tam, A., Mcclelland, J., ... & Hill, F. (2022, June). Tell me why! explanations support learning relational and causal structure. In *International conference on machine learning* (pp. 11868-11890).

[80] Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L.,… & Batson, J. (2025). On the biology of a Large Language Model. *Anthropic Research*.

[81] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

[82] Xiang, V., Snell, C., Gandhi, K., Albalak, A., Singh, A., Blagden, C., ... & Finn, C. (2025). Towards system 2 reasoning in LLMs: Learning how to think with meta chain-of-thought. *arXiv preprint arXiv:2501.04682*.

[83] Porter, T., Elnakouri, A., Meyers, E. A., Shibayama, T., Jayawickreme, E., & Grossmann, I. (2022). Predictors and consequences of intellectual humility. *Nature Reviews Psychology*, *1*(9), 524–536.

[84] Basseches, M. (1980). Dialectical schemata: A framework for the empirical study of the development of dialectical thinking. *Human Development*, *23*, 400-421

[85] Ritter, F. E., Tehranchi, F., & Oury, J. D. (2019). ACT-R: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*, e1488.

[86] Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, *33*, 1–64.

[87] Goertzel, B., Bogdanov, V., Duncan, M., Duong, D., Goertzel, Z., Horlings, J., ... & Werko, R. (2023). OpenCog Hyperon: A framework for AGI at the human level and beyond. *arXiv preprint arXiv:2310.18318*.

[88] Vezhnevets, A. S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., & Kavukcuoglu, K. (2017). FeUdal networks for hierarchical reinforcement learning. In *International conference on machine learning* (pp. 3540–3549).

[89] Zhang, K., Yang, Z., & Başar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. In K. G. Vamvoudakis et al. (Eds.), *Handbook of reinforcement learning and control* (pp. 321–384).

[90] Hassan, S., & De Filippi, P. (2021). Decentralized autonomous organization. *Internet Policy Review*, *10*.

[91] Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. *arXiv preprint arXiv:1805.00899*.

[92] Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

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 to 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 provide conflicting advice.
- **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.
- **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 heuristic, narrative, or analytic procedure.
- **outcome-monitoring process:** A type of metacognitive process that checks whether outcomes of the selected object-level strategy are plausible (also called "sanity checking").
- **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).