Choosing the Path of Most Resistance: Metacognitive AI Literacy and Mitigating Epistemic Risk

Iman Khwaja
University of Edinburgh

This paper reimagines metacognition as both an individual tool and a component of AI (Artificial Intelligence) literacy, that acts as a critical pathway to mitigating the epistemic threats posed by AI tools. These threats are explored as the erosion of autonomy, critical thought and creative abilities in individuals, as a result of cognitive dependencies on AI tools.

Metacognition, a strategy involving the active preservation of cognitive ability in individuals, is currently a subtle component of many literacy approaches in varied educational contexts. This paper positions it as a practice that could play a meaningful role in the trajectory of AI development in tandem with the consequent changes to our collective epistemic and cognitive experiences. Specifically, the aspects of self-restraint, critical evaluation and epistemic investigation proposed by traditional metacognitive strategy are reconceptualized as potentially critical aspects of AI literacy efforts.

Drawing on literature that explores AI deskilling, epistemic risk, which Babui describes as "the likelihood that one's claims inaccurately represent the world" (2019), and digital literacy, this paper proposes a framework for users, developers and deployers to consider applying metacognitive strategy to AI literacy efforts, as a means to mitigate the epistemic risks posed by user dependencies on AI tools.

Keywords: AI Literacy, Metacognition, Epistemia, Epistemic Risk, AI Ethics

Introduction

The proliferation of Artificial Intelligence (AI) models is changing the landscape of literacy and knowledge acquisition in an unprecedented way. AI models that operate by scraping and synthesizing large amounts of data to produce personalized responses to user queries are referred to in this paper generally as "digital concierges", as an umbrella term for LLM's (large language models) and other models that operate similarly. These responses have been described as "oracular" (Wihbey, 2024) in nature, in which the black-boxing of the data behind these informational outputs and the elimination of the traditional search by search engine process allows users to avoid tracing information to its source and evaluating its validity. While these rapid and articulate outputs offer evidently helpful enhancements to our own epistemic experiences, the question remains as to whether we are able to protect our cognitive wellbeing amidst increasing epistemic accessibility. This paper advocates for the development of literacy strategies that acknowledge the cognitive implications of AI use in tandem with its development. Southworth et al. reaffirm this suggesting that "*academic and industrial bodies need, in a symbiotic way, to reaffirm responsible behaviour*" (2023), which this paper proposes can be done through embedding reconceptualized metacognitive strategies to support users in various use contexts.

This paper poses the following questions as a means to investigate these opportunities preliminarily, and to prompt further research into metacognition as a component of decision-making for educators, users and developers of AI technologies interested in cognitive wellbeing:

1. Should cognitive wellbeing be a critical consideration in future AI literacy efforts in the context of epistemic AI use?

2. How can existing metacognitive frameworks be reimagined and applied to encourage individual and broader regulation of cognitive wellbeing?

3. What adjustments can and should institutions make to augment the

responsible and critical epistemic uses of AI?

4. How could a broad metacognitive approach to AI use encourage appropriate reliance on AI tools across a variety of use cases?

This paper identifies several key concepts and preliminarily outlines these below, that form the basis of the response this paper constructs to the above questions.

1. **Metacognition**: The ability to work critically with information and conduct reasoning in response to information by evaluating it using epistemic beliefs. Pertaining to knowledge about human cognition and the regulation of these cognitive processes with systemic, institutional and self-directed support.

2. **Epistemic Risk**: Generally threats to the ability to discern validity, truth and critically evaluate knowledge, the sources of knowledge and the structures of knowledge with accuracy.

3. **Cognitive Wellbeing**: The status of judgement, reasoning, critical thinking and assessment capabilities in individuals. With atrophy of these facets of cognition being a limitation to this wellbeing, being able to employ these in good standing in response to information is indicative of cognitive wellbeing (Salovich et al, 2021).

4. **Appropriate Reliance**: John Coleman's argument that "*AI is already beginning to sit between people and their own thinking - how we read, write, reason and form beliefs*" provides a starting point to determine appropriate reliance. This concept broadly refers to how users are able to contextualize their use of AI in a way that does not impede their cognitive wellbeing or pose an epistemic risk to their thinking processes. Contextual use is paramount in consideration of appropriate reliance, where some users have more autonomy over determining their use of and exposure to AI tools than others depending on their required application.

Brief Assessment of Current AI Literacy Efforts

The current landscape of AI literacy is arguably limited. Amongst the presence of many AI ethics and adjusted digital skills courses available to individuals online, there is a clear lack of standardised AI literacy programs as enforced components of national education in the UK. The publication of the AI Opportunities Action Plan in early 2025 written by Matt Clifford recalls a need to "*support institutions to increase the numbers of AI graduates and teach industry-relevant skills*" (p. 9) but does not detail how these graduates and students will be supported to be well-adjusted and cognitively strong users of these technologies.

The creation of dependence is a concern expressed across the literature, summarised by Kalantzis' dilemma: "*Why read, when a text can be spoken to you, and can you be told to speak with just the right level of difficulty for you?*" (2024). In the last year, numerous efforts to facilitate an ethical and critical approach to the perception and use of AI have emerged. In 2023, Kanta Dihal and Tania Duarte published Better Images of AI: A Guide for Users and Creators, as a primary reference point to facilitate "*accurate and compelling communication*" about AI (p. 3) in visual contexts. Similar conceptualizations such as The Philosophical Glossary for AI, created by Dr Alex Grzankowski for the London Humanity Project, consist of entries by various authors working to redefine existing terms such as "consciousness" and "bias" in the context of relationality to AI tools, that reveal the critical importance of reconceptualization in the AI literacy process itself. Interestingly, Brian Ordegaard and Diego Morales submitted an entry to the glossary redefining "metacognition" as a capability held by AI systems themselves.

While this paper acknowledges the value of exploring this iteration of metacognitive ability in digital contexts, this paper goes one step further to argue that employing metacognitive strategy as a responsive measure to the use of AI systems is a critical discussion that should coexist with exploring parallel abilities in AI models. In either case, the definition of metacognition as "*a means for systems to assess their own reliability*" that Ordegaard and Morales have outlined is useful to the arguments presented in this paper. This "assessment" process is therefore the central premise of the metacognitive framework constructed in this paper. Within this, the abilities to "*remember, reason, judge, think, and believe*" are included as cognitive processes that constitute metacognitive practice.

For the sake of discussion, this paper uses the phrases "knowing thyself" and "choosing the path of most resistance" to informally summarize the complexity of metacognition, a reiteration of what Yi describes as "knowing what to know" in their definition of metacognition (2025). Importantly, Yi also describes "the purpose of AI literacy" as "to have anticipation capabilities", in which "*metacognition involves an effort to anticipate an uncertain future*" (2025). These definitions of metacognition as a constituent competency of broader AI literacy are pertinent to the discussions outlined in this paper, which will be summarized in a framework that hybridizes traditional metacognitive concepts and their reconceptualizations for applied AI literacy strategy.

This paper positions this framework as an interventional strategy that ensures that "*the convenience offered by AI tools does not come at the cost of essential cognitive skills*" (Zhai et al, 2024). In "choosing the path of most resistance", I invite users, developers and educational deployers to engage in and promote metacognitive practice as a means to preserve their epistemic autonomy.

1. What is metacognition?

Developmental psychologist John Flavell coined the term metacognition in 1976 to denote the concept of "cognizing cognition itself" (Noushad, 2008). While definitions are varied and the historical roots of metacognition are complex, this paper benefits from the qualities of "self-appraisal" and "self-consciousness" that (Noushad, 2008) suggest are fundamental to metacognition, reiterated by Flavell's own assertion that "*metacognitive strategies monitor the process of learning and task completion*" (1976). Baker and Brown's assertion that Flavell's idea consisted of two main "*clusters of activities: knowledge about cognition and regulation of cognition*" (1980) suggest the importance of self-intervention across the board, getting to know one's own patterns of cognition and acquiring the skills to regulate them through continuously developed familiarity with these patterns, thus centralising an awareness of the self as crucial for effective metacognitive practice.

External to Flavell's view that "*critical thinking is subsumed by metacognition*" (1979), central ideas of self-regulation and monitoring are present in a wide spectrum of definitions of metacognition. Hennessy demonstrates this in their description of the "*multidimensional character of the term metacognition as: (1) an awareness of one's own thinking; (2) an awareness of the content of one's conceptions; (3) an active monitoring of one's cognitive processes*" (1999). This definition is consistent with Noushad's analysis of the "self-regulatory mechanisms" involved with metacognition, "*checking outcomes, planning the next move, monitoring effectiveness, testing, revising and evaluating one's strategies for learning*" (2008).

Consequently, various sources suggest metacognition is akin to critical thinking in many respects and could therefore be defined concretely as: the awareness of one's own cognitive processes and the continuous self-regulation and monitoring of these cognitions through an intentional and critical revisiting of them. To return to Yumi's definition of metacognition for AI literacy specifically, "knowing what to know" and "the anticipation of ambiguity" affirm the essence of these early definitions of metacognition.

Contemporary metacognitive research remains largely in the realm of its applications to traditional education. However, this paper attempts to explore the role of metacognition in

the context of digital epistemologies. It is therefore important to align the shifting landscape of epistemologies with the relevant principles of metacognition, to ensure it remains an applicable and useful practice in the face of unprecedented technological development. The value of metacognition in mitigating epistemic risk and preserving cognitive wellbeing against inaccuracy is relatively well recorded in the body of literature, where Salovich et al. conducted an experiment investigating this that applied metacognitive tasks to encourage evaluation in participants, demonstrating that participants "*made fewer judgement errors after reading inaccurate statements*" after completing it. This paper argues that a similar, adjusted practice could protect users against the implications of AI generated inaccuracies. This calls for a basic referential framework for the principles of metacognition to be reconceptualized, so that they may be ascertained and adjusted according to their utility for engagement with AI.

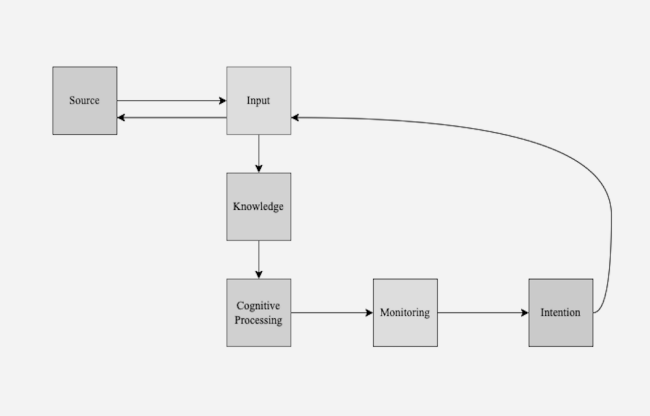

Figure 1: Referential Model for Metacognitive Practice

1.1. Decoding Metacognition: Figure 1 Key

Source: Refers to the source of information itself in any capacity, such as visual or textual inputs.

Input: Refers to how the learner goes about acquiring the relevant knowledge required to perform the task, such as locating the information spatially, asking a question or inputting search terms manually or digitally. The bidirectional arrow between source and input is to demonstrate the relationship between the existence of the source in its entirety and the learner's intentionality to locate relevant information within this greater whole.

Knowledge: This term refers to the process that takes place prior to the extraction of meaning from the input or any inference that takes place and remains an abstraction or a conceptualization based on a priori understandings of the input.

Cognitive Processing: This term broadly encompasses the information ascertained by the learner. This involves the processes of "*remembering, reasoning, representing, thinking and believing*" outlined by Ordegaard and Morales in their definition of metacognition.

Monitoring: The distinctive aspect of metacognition, this stage is the crucial evaluation, reflection and other "self-regulatory mechanisms" (Baker and Brown, 1999) that allow the learner to retrace the steps of their own cognitive processing and knowledge stages. This is the formation of awareness of how this input was received by the learner and where repetitive learning patterns are established that allow the learner to make retroactive changes to cognitive processing.

Intention: I have added this to the diagram despite it not being explicitly mentioned in the metacognition body of literature. In this visual it acts as a consequential pathway that demonstrates how metacognitive activity changes the learner's intentions to pursue information in ways that are familiar, effective, or continually ineffective, depending on the success of the monitoring stage. It acts as a strategic factor that is a result of the "re-visiting" process in metacognition and determines how the learner plans to approach future situations.

1.2. Defining Epistemic Threat

Threats emerging from the use of intelligent technology are numerous and ever-changing. Epistemic threat considerations currently remain in the background of most large-scale risk mitigation efforts, but it is not difficult to visualize how AI tools that take on "digital concierge" roles by providing users with personalised epistemic experiences could become threatening in their own way. Crucially,

identifying these threats is part of what Yi describes as the importance of using literacy to support our collective abilities to "*explore uncertain and complex societies, and predict future problems and find solutions*" (2021). LaGrandeur subsequently defines epistemic risk as a threat to our capacities to "*make decisions, conduct analytical reasoning and regulate our creative urges*" (2021), which this paper identifies as the components of metacognition threatened by epistemic dependence on AI proxies. Faced with resolving these dependencies, a widely applicable literacy strategy in tandem with the self-regulation aspects of metacognition is proposed in this paper as a pathway towards mitigating at least some of the epistemic threats posed by proliferating complex technologies.

Several empirical cases of epistemic risk have emerged in a growing body of evidence demonstrating an increase in concerning user-machine relationships. Kosymna et al. conducted a study producing neurological evidence through electroencephalography (EEG) scanning for the phenomenon of "cognitive debt" emerging in participants, where the results showed "*diminished critical thinking, reduced creativity and independent thought, vulnerability to manipulation and shallow information processing*" in groups that delegated tasks to A (2025) .

Further, a report by Baracui evaluating AI as a supplement for academic study indicated that the unrestricted use of AI tools in undergraduate students impeded retention capabilities for academic materials (2024). These findings not only indicate an exacerbation of cognitive complications in generalized AI use, but also shifts in the perceptions of these agents as they become increasingly personalized. The extremes of emotional dependency on AI models are now well documented, particularly in the wrongful death lawsuit filed by the family of sixteen-year-old Adam Raine against OpenAI in August 2025, that claimed OpenAI's leading chatbot ChatGPT played a role in his decision to take his own life. Whether the failure exhibited by the model was technical or an issue of guardrail development, the epistemic complications of Raines' relationship with ChatGPT as an agent itself remains of interest.

Within the scope of this paper, epistemic threat is considered as related to complications of dependency, loss of critical thought, clouded judgement and perception and the general erosion of cognitive wellbeing. This paper briefly identifies the following hypothetical epistemic threats with the intention to contribute to mitigating them. The threats identified use examples that position the end-user in an affected position, to illustrate the potential for epistemic threats to proliferate without intervention. In identifying these epistemic threats, the intention is not to be presumptuous about the realism or probability of these outcomes, but to depict them in a hyperbolised manner, to prompt further research on interventional response measures that cover the full extent of possibility.

Dependence: The establishment of a reciprocal dependency between the user and the AI algorithm where increasingly personalised outputs replace the user's ability to engage with other neutral and non-personalised sources of information. The user therefore is unprepared to engage with sources that do not produce outputs that reflect the users' cognitive patterns as ascertained by the algorithm. The user depends on the AI as a singular source for most tasks, damaging the user's relationship with other sources of information.

Cognitive Offloading: A by-product of dependence, the user becomes comfortable offloading cognitive processes such as critical thinking, evaluation, and judgement to the AI tool. This might look like asking the AI to critically evaluate a source on their behalf and adopting this evaluation as their own without conducting an objective appraisal of it, without interference or consultation from the AI. The user is not able to break the barrier of superficial understanding of key concepts due to the rapid generation of seemingly accurate and useful information.

Loss of Decisive Autonomy: The user's ability to make decisions about how to approach educational tasks is damaged significantly, and the user struggles to trust their own decision-making skills without the guidance or advice of an AI tool that algorithmically reflects their interests. The user is uncomfortable engaging

with cognitive trial and error to ascertain the best strategies for their own learning outcomes.

Erosion of Critical Thought and Judgement Capabilities: The user becomes unable to critically evaluate sources and form their own judgements without consulting the AI to seek confirmation, affirmation, or guidance. The user is unable to organically assess sources against other textual or digital sources without requiring that they be summarised in an accessible and personalised format.

Creative Injury: The user struggles to extract meaning from sources and produce novel or original solutions to challenges. The user loses the ability to adapt to new styles, formats, or contexts of information, and equates creative and experimental outputs with threatening unfamiliarity. The user no longer perceives a relationship between emotional resonance and creative cognition and perceives creativity as possible only through digital mediums or in digital contexts. The user becomes unable to adapt to new, unfamiliar situations and is discomforted by the challenge of creative problem-solving with other humans to meet demands. The user becomes comforted by the confirmation bias offered by the AI tool and no longer seeks collaborative creative work with other humans.

Intuitive Confusion: Users struggle to make intuitive decisions about when an AI tool should be overridden in decision-making instances. Users struggle to reconcile their a priori knowledge and beliefs with the recommendations of the AI and become intuitively confused about actions thereafter. Users are impacted intuitively and are unable to make intuitive decisions or judgements without AI assistance or confirmation.

Perceptual Fragmentation: The user no longer perceives the tool as a recursive product of their own algorithmic interactions, they become familiar with the tool as an extension of their own cognition or perceive it as a cognitive tool to fill gaps in their own capabilities. The user can no longer perceive their own cognitions, judgements, and independent criticisms as unique or separate. The user perceives AI as intellectually and creatively superior and therefore shifts their perception of their own capabilities and the capabilities of others.

The above threats have not been concretely identified in any official capacity but are present across the literature as common concerns about the reliance on AI modalities for educational and personal information. In establishing these epistemic threats in a more concrete manner, we create opportunity pathways for response capabilities. The following section will explore how metacognitive principles can be adjusted for AI to represent these protective measures in a widely applicable and referential way.

2. How Can Metacognition be Adjusted to Protect Against Epistemic Atrophy Caused by AI?

In response to the above outlines of metacognitive principles and epistemic threat, a conceptualization of actionable metacognitive markers can be developed in the context of AI use.

The applicability of this is to create an interventional starting point for researchers, deployers and educators to reference as a means to embed risk awareness and the natural preservation of cognitive wellbeing into AI literacy and epistemic risk mitigation efforts alike. The target of cultivating metacognitive thinking across literacy disciplines and in individuals themselves is ultimately to encourage users to continually and critically examine their own abilities to evaluate AI generated information.

Primarily, this means gearing metacognitive thinking towards recognizing what Wihbey describes as the "*incompleteness of AI's judgement with respect to human knowledge and values*" (2024), and therefore an awareness of appropriate reliance on these technologies that is contextual. This is a logistically challenging assertion in that reliance and dependency can be largely circumstantial depending on the users fluctuating needs, environment and requirements to use AI models, and this paper recognizes that despite attempts to categorize epistemic threat and metacognitive practice, not all users benefit from this proposed framework.

To illustrate this limitation, it should be considered that a student using an AI tool for academic purposes, such as to synthesize a body of literature for research purposes and to consult with it for academic support would be more of an epistemic risk consideration than a radiologist consulting with an AI medical imaging tool for detection purposes. However, this does not exempt the radiologist from the benefit of metacognitive practice in that they retain an awareness of their oversight value and contextual medical knowledge as a physician when evaluating information outputs.

Within these contexts we can achieve "*a development of students' metacognitive skills to help them become aware of when and how to use AI tools appropriately without undermining their cognitive development*" (Gerlich, 2025). This implies that "appropriate reliance" (Kong et al, 2021) contexts for our engagement must be differentiated from the outset, to develop principles of appropriateness that allow end-users to review their motives for engaging with AI in epistemic contexts. Here we have the first few components of what metacognition for AI literacy could look like, an expansion of the traditional self-reviewing of learning behaviours to a comprehensive review of our own awareness of multiple factors that impact our cognitive health, such as knowledge of algorithmic infrastructures, changing policies and emerging threats.

Additionally, an understanding of the limitations of AI design, and indeed the "discernible cause and effect" (Zittrain, 2022) that produces its judgements, and how this is vitally different from our own. A metacognitive framework that achieves these various objectives should therefore be widely applicable and should be referential in both "classroom ecologies" (Kalantzis et al, 2024) and instances of individual need alike. This argues for the retention of what is essential to our cognitive health whilst leaving room for changes to our cognitive expenditure. Kuhn and Dean reiterate the value of metacognition in their explanation that "*metacognition is what enables a student who has been taught a particular strategy in a particular problem context to retrieve and deploy that strategy in a similar but new context*" (2004), this implies that a fundamental metacognitive framework could allow for wide cross-contextual use.

For this reason, I argue that the overarching principles of self-review, self-restraint and continuous self-assessment can be merged into the ability to simply "know thyself", from which the tenets of intention, critical evaluation and contextualised judgement emerge thereafter.

Originating from an inscription on the Temple of Apollo at Delphi, "know thyself" is a Greek maxim that denotes the importance of self-awareness principally (Green, 2024). In a literacy setting, this is not equivalent to personal introspection or philosophical musings on consciousness but is simply the idea that end-users must be empowered by and made aware of their own cognitive, critical, and creative capabilities, before engaging with seemingly equally intelligent and advanced technologies. Thus, a metacognitive framework revolving around this principle will not only highlight these functions but emphasise the importance of preserving them and tending to them against the impulse for convenience and ease.

This second main component is what I have referred to as "*choosing the path of most resistance*". The framework should emphasise the importance of doing so circumstantially, be it preserving our investigative skills by recommending self-directed search strategies or encouraging dialogue with collaborators, before engaging with AI tools that are "properly calibrated for educational application" (Kalantzis et al, 2024). This means end-users should ultimately consider on a circumstantial basis how their use of AI weighs against their awareness of their cognitive health and evolving epistemic risks and attempt to reconcile these by "choosing the path of most resistance" in an actionable way. Barbara Grosz's development of the "Intelligent Systems: Design and Ethical Challenges" Course for the Harvard University Department of Computer Science combines these concepts in a literacy context, emphasizing that they wish to "*educate students to think not only about what systems they could build, but whether they should build those systems and how they should design them*" (2019). These interdisciplinary moves toward

educating students about their own intentions, motives, and engagement with AI highlights what a metacognition framework for AI literacy could offer developers and end-users alike. The following section presents this framework as a comprehensive and referential model.

2.1. Epistemic Modesty: Performance Limitations and Appropriate Reliance

In a Substack post for the Cosmos Institute written by Professor Kevin Vallier titled On the Noble Uses of AI, Vallier argues for the "*inevitability of cognitive atrophy*" amidst epistemic augmentation. Using historical examples of our deference to calculators to perform arithmetic function, Vallier's argument is closely aligned with the concept of appropriate reliance reiterated in this paper (2026). The possibilities for AI tools to enhance our intelligence, performance and overall capabilities are only growing in tandem with development efforts. Vallier presents these opportunities to cognitively offload reductive tasks as a means to "focus on other things", where "hunting for sources" when conducting research allows researchers to instead think about "what the sources mean" (2026).

It is the delicate balance between succeeding at rejecting AI's attempts to "*seduce us into passivity*" (Vallier, 2026) and to cognitively defer to these tools as a means to prioritize more meaningful work that this paper terms as "appropriate reliance", and argues that metacognitive practice adjusted to meet this challenge can help be cultivated in end users. This preservation of deliberation, reasoning and judgement capacities by investigating information synthesized by AI is the essence of what metacognitive practice applied to AI-driven contexts can achieve, but it is equally crucial that these skills are preserved alongside our knowledge of the epistemic modesty of these tools, as well as the epistemic risks posed to us by our dependencies on them.

There are several cases of AI models producing epistemically unreliable outputs, in a study conducted by Urban et al, findings revealed that 88% of students who relied on ChatGPT to supplement processes of academic writing demonstrated an enhanced performance, but that "*their own thinking remains crucial to effectively integrating information from human-generated sources not included in the training dataset of generative AI tools*" (2024). Hannigan et al. reiterate this in their argument that "*response veracity is crucial*" when assessing AI generated information, and suggests that users "*likely need to employ critical thinking skills, inductive reasoning and forecasting to estimate the veracity of a statement produced by a chatbot*" (2024).

These indications of combined cognitive augmentation and confined reliance on these tools suggest the need for referential literacy frameworks that support users in determining information veracity and achieving appropriate reliance. The following sections propose an informative model for further research and educators as a means to consider embedding this into AI literacy curriculums.

2.2. Metacognition for AI: Actionable Framework

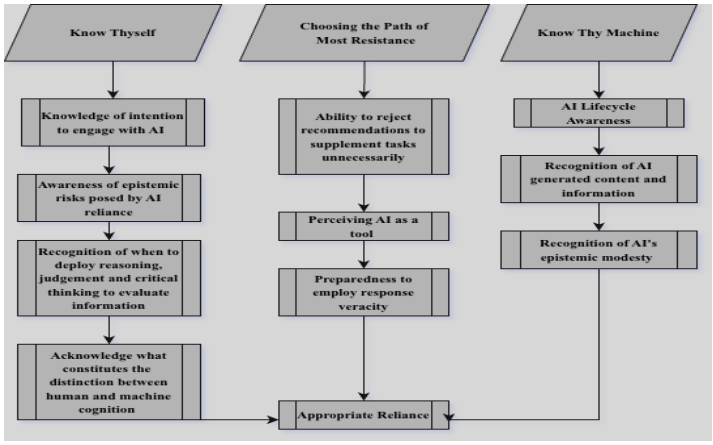

Figure 2: Referential Model for Metacognitive Practice for Applied AI

The below framework provides a visual summary of the content of this paper's proposition for an adjusted metacognitive application for AI use. The framework is intended to outline conceptual principles that ultimately constitute appropriate reliance, which can be expanded to various use contexts. The framework is underpinned by hybridizing risk awareness and knowledge of both AI and human cognition to determine appropriate use of AI tools for epistemic purposes.

2.3. Decoding Metacognition for AI

The following section investigates the concepts outlined in Figure 2 in further depth to provide

clarification on what the metacognitive process itself entails.

Know Thyself: Component of awareness in the metacognitive process that preserves user autonomy and makes a crucial distinction between themselves and their perceived machine counterparts. Empowering users to embrace their uniquely human cognitive and epistemic experiences preceding their interactions with AI involves alerting users to the basic operations of human cognition and machine cognition, to make crucial distinctions between them known.

Know Thy Machine: The cultivated capability to criticize new kinds of information is a skill that this paper argues AI literacy curriculums should begin to cultivate in users who experience epistemic dependence on AI in any capacity, across disciplines and applications. This includes creating an awareness of the AI lifecycle of development, the role of data in exacerbating and embedding bias into AI generated information and overall working to demystify the operations of AI tools by revealing the basics of their processing operations to users. Akin to the "knowledge" aspect of the metacognitive framework presented earlier in this paper, "knowing thy machine" is critical to developing cognitively and epistemically sound relationships with these tools by creating a fundamental awareness of their limitations.

The Path of Most Resistance: This practice expands the metacognitive practices of revisiting, monitoring and investigating information itself to describe being equipped to critically evaluate AI outputs. In recognizing the epistemic modesty of AI tools that is often black-boxed behind the personalized and highly accessible nature of their outputs, users have the opportunity to critically evaluate and contextualize outputs by enacting reformed metacognitive skills, such as critical reasoning and evaluative judgement, to decide the validity of these outputs. Specifically, the ability to reject recommendations offered by AI tools designed to supplement work, such as to perform administrative tasks or to rewrite academic content for students, is an important capacity that users have that they must at least retain an awareness of. With the knowledge of technical fallibility, embedded bias and the inner workings of AI models themselves, users can determine for themselves, or learn to determine with the support of literacy frameworks, how to engage with models appropriately, and when to actively choose making a cognitive effort to protect themselves from epistemic risk.

3. Limitations

This paper does not explore metacognitive practice in enough depth to produce a robust, scalable, and integrable solution for its adjustment to AI solutions. It anticipates successful adherence to frameworks without considering the autonomous decisions individuals take to use AI to optimise their own outcomes. Developing valid and reliable assessment tools for metacognition is also a challenge, as is the ability for individuals to conduct these assessments on themselves. The application of such a framework is also limited by epistemic variability, and varied epistemic norms in global digital curricula. This framework also does not account for heuristic needs and cannot produce reliable results in all learning scenarios. Much of the success of this cognitive protection is reliant on individual willingness, ability and circumstance, and this paper recognises the need for further research to support how other standardised and quantifiable solutions can be developed.

Conclusion

This paper presents a summary of findings that position metacognition as a protective, interventional strategy to mitigate epistemic and cognitive threats posed by AI generated information. Recognizing that the structure of knowledge itself has changed unprecedentedly with the advent of AI, adjusting traditional architecture for praxis in new contexts requires identifying these changes and the subsequent emerging threats from them. The implications for this kind of research are numerous but remain in the preliminary stages of application.

The primary complications highlighted in this paper are epistemic threats and distortions of our responses to information, as a result of dependencies on AI technologies. The framework proposes applicable but conceptual ideas that are intended to act as a reference point for educators, users and developers

dealing with exposure to these tools to support their epistemic autonomy and cognitive wellbeing. Ideally, further research that takes on a more empirical approach to actioning these concepts in various use contexts would legitimize these early investigations.

Conclusively, this paper advocates for meaningful epistemic interactions with AI technologies that are actively governing how we think and who we become.

References

Baker, L. and Brown, A.L. (1984) 'Metacognitive Skills and Reading', in Pearson, P.D., Barr, R., Kamil, M.L. and Mosenthal, P. (eds.) *Handbook of Reading Research*. New York: Longman, pp. 353–394.

Babic, Boris. "A Theory of Epistemic Risk." *Philosophy of Science* 86.3 (2019): 522-550.

Biswas, M. and Murray, J. (2025) '"Incomplete Without Tech": Emotional Responses and the Psychology of AI Reliance', in Huda, M.N., Wang, M., and Kalganova, T. (eds.) Towards Autonomous Robotic Systems. TAROS 2024. *Lecture Notes in Computer Science*, vol. 15051.

Biswas, M. and Murray, J. (2024) 'The Impact of Education Level on AI Reliance, Habit Formation, and Usage', 2024 29th *International Conference on Automation and Computing* (ICAC), Sunderland, United Kingdom, pp. 1–6.

Borenstein, J. and Howard, A. (2021) 'Emerging Challenges in AI and the Need for AI Ethics Education', *AI and Ethics*, 1(1), pp. 61–65.

Carr, N. (2010) 'The Shallows: What the Internet is Doing to Our Brains', *Synthese*, 178(3), pp. 341–348.

Cardon, P., Fleischmann, C., Aritz, J., Logemann, M., & Heidewald, J. (2023) 'The Challenges and Opportunities of AI-Assisted Writing: Developing AI Literacy for the AI Age', *Business and Professional Communication Quarterly*, 86(3), pp. 257-295.

Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B., & Barro, S. (2023). AI Literacy in K-12: A Systematic Literature Review. *International Journal of STEM Education*, 10(1), Article 29.

Chen, V., Liao, Q.V., Vaughan, J.W., & Bansal, G. (2023) 'Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations', *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2), Article 370.

Chein, J.M., Martinez, S.A., & Barone, A.R. (2024) 'Human Intelligence Can Safeguard Against Artificial Intelligence: Individual Differences in the Discernment of Human From AI Texts', *Scientific Reports*, 14, Article 25989.

Chiu, T.K.F., Ahmad, Z., Ismagilov, M., and Sanusi, I.T. (2024) 'What Are Artificial Intelligence Literacy and Competency? A Comprehensive Framework to Support Them', *Computers and Education Open*, 6, 100171.

Department for Science, Innovation and Technology. (2025). AI Opportunities Action Plan. London: Department for Science, Innovation and Technology.

Dihal, Kanta, and Tania Duarte. Better Images of AI: A Guide for Users and Creators. The Leverhulme Centre for the Future of Intelligence / We and AI, Feb. 2023, blog.betterimagesofai.org/wp-content/uploads/2023/02/Better-Images-of-AI-Guide-Feb-23.pdf.

Flavell, J.H. (1979) 'Cognitions About Cognitions: The Theory of Metacognition', *American Psychologist*, 34(10), pp. 906–911.

Gerlich, M. (2025) 'AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking', SSRN Electronic Journal.

Metacognition." *The Philosophical Glossary for AI*, edited by Alex Grzankowski and Benjamin Henke, 18 Dec. 2025, aiglossary.co.uk/2025/12/18/metacognition/.

Heersmink, R. The Internet, Cognitive Enhancement, and the Values of Cognition. *Minds & Machines 26*, 389–407 (2016).

Holstein, K., Aleven, V., McLaren, B. M., & Sewall, J. (2020). Addressing Student Needs with Human and Artificial Intelligence Tutors. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 1982–1991.

Jobin, A., Ienca, M., & Vayena, E. (2019) 'The Global Landscape of AI Ethics Guidelines', *Nature Machine Intelligence*, 1(9), pp. 389–399.

Johnson, B. (2022) 'Metacognition for Artificial Intelligence System Safety – An Approach to Safe and Desired Behavior', *Safety Science*, 151, 105743.

Kalantzis, M. and Cope, B. (2025) 'Literacy in the Time of Artificial Intelligence', *Reading Research Quarterly*, 60(1), Article e591.

Karoff, P. (2019) 'Embedding Ethics in Computer Science Curriculum', *Harvard Gazette*, 25 January. Available at: https://news.harvard.edu/gazette/story/2019/01/harvard-works-to-embed-ethics-in-computer-science-curriculum/ (Accessed: 9 February 2025).

Keles, S. (2023) 'Navigating in the Moral Landscape: Analysing Bias and Discrimination in AI through Philosophical Inquiry', *AI and Ethics*, 3(4), pp. 901–917.

Kong, S.-C., Cheung, W.M.-Y., & Zhang, G. (2021). Evaluation of an Artificial Intelligence Literacy Course for University Students with Diverse Study Backgrounds. *Computers and Education: Artificial Intelligence*, 2, 100002.

Kuhn, D. and Dean, D. (2004) 'A Bridge Between Cognitive Psychology and Educational Practice', *Theory into Practice*, 43(4), pp. 268–273.

LaGrandeur, K. (2020) 'How Safe is Our Reliance on AI, and Should We Regulate It?', *AI and Ethics*, 1(1), pp. 67–73.

Liu, Jiahui. 2025. "When Usefulness Fuels Fear: The Paradox of Generative AI Dependence and the Mitigating Role of AI Literacy." *International Journal of Human–Computer Interaction*, August, 1–22. doi:10.1080/10447318.2025.2544006.

Luccioni, A. and Bengio, Y. (2020) 'On the Morality of Artificial Intelligence', IEEE *Technology and Society Magazine*, 39(4), pp. 36–43. Available at: https://web.archive.org/web/20201108110312id_/https://ieeexplore.ieee.org/ielx7/44/9035512/09035523.pdf (Accessed: 9 February 2025).

Ng, D.T.K., Leung, J.K.L., Chu, K.W.S., & Qiao, M.S. (2021). AI literacy: Definition, Teaching, Evaluation, and Ethical Issues. *Proceedings of the Association for Information Science and Technology*, 58(1), 504–509.

Pritchard, D. (2010) 'Cognitive Ability and the Extended Cognition Thesis', *Synthese*, 175(1), pp. 133–151.

Przybylski, A.K., Murayama, K., DeHaan, C.R., & Gladwell, V. (2013) 'Motivational, Emotional, and Behavioural Correlates of Fear of Missing Out', *Computers in Human* Behavior, 29(4), pp. 1841–1848.

Salovich, N. A., & Rapp, D. N. (2021). Misinformed and Unaware? Metacognition and the Influence of Inaccurate Information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(4), 608–624.

Serafimova, S. (2020) 'Whose Morality? Which Rationality? Challenging Artificial Intelligence as a Remedy for the Lack of Moral Enhancement', *Humanities and Social Sciences Communications*, 7, Article 119.

Sparrow, B., Liu, J. and Wegner, D.M. (2011) 'Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips', *Science*, 333(6043), pp. 776–778.

Southworth, J., Migliaccio, K., Glover, J., Glover, J., Reed, D., McCarty, C., Brendemuhl, J., & Thomas, A. (2023) 'Developing a Model for AI Across the Curriculum: Transforming the Higher Education Landscape via Innovation in AI Literacy', *Computers and Education: Artificial Intelligence*, 4, 100127.

Tock, J.L. and Moxley, J.H. (2017) 'Metacognition: A Predictor of Learning

Outcome', *The Journal of Educational Research*, 110(5), pp. 478–489.

Touretzky, D., Gardner-McCune, C., Breazeal, C., Martin, F., & Seehorn, D. (2019). A Year in K-12 AI Education. *AI Magazine*, 40(4), 88–90.

UK Government (2025) International AI Safety Report 2025. Available at: https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf (Accessed: 9 February 2025).

Urban, Marek, et al. ""ChatGPT Can Make Mistakes. Check Important Info." Epistemic Beliefs and Metacognitive Accuracy in Students' Integration of ChatGPT Content into Academic Writing." *Computers & Education*, vol. 225, 2025, p. 105118. ScienceDirect, https://doi.org/10.1016/j.compedu.2024.105118.

Vartiainen, H., Toivonen, T., Jormanainen, I., Kahila, J., Tedre, M., & Valtonen, T. (2021) 'Machine Learning for Middle Schoolers: Learning through Data-Driven Design', *International Journal of Child-Computer Interaction*, 27, 100281.

"Vallier, K., On the Noble Uses of AI: When Cognitive Offloading Elevates Us." *Cosmos Institute*, 30 Jan. 2026, blog.cosmos-institute.org/p/on-the-noble-uses-of-ai.

Voeneky, S., Kellmeyer, P., Mueller, O., and Burgard, W. (eds.) (2022) *The Cambridge Handbook of Responsible Artificial Intelligence*. Intellectual Debt, Zittrain, J, Chapter 11. Cambridge: Cambridge University Press.

Wilton, L., MacKinnon, K., & Greene, K. (2021) 'Where is the AI? AI Literacy for Educators', in Artificial Intelligence in Education. AIED 2021. *Lecture Notes in Computer Science*, vol 12749. Springer, Cham, pp. 137–142.

Wihbey, J. (2024) 'AI and Epistemic Risk for Democracy: A Coming Crisis of Public Knowledge?', *SSRN Electronic Journal*.

Yi, Y. (2021). Establishing the Concept of AI Literacy. Jahr – *European Journal of Bioethics*, 12(2).

Zohar, A. and Barzilai, S. (2013) 'Promoting Self-Regulation in Science Education: Metacognition as Part of a Broader Perspective on Learning', *Metacognition and Learning*, 8(1), pp. 99–111.

Zhai, C., Wibowo, S. and Li, L.D. (2024) 'The Effects of Over-Reliance on AI Dialogue Systems on Students' Cognitive Abilities: A Systematic Review', *Smart Learning Environments*, 11, Article 28.