Proposed *Nature* Perspective

Cognitive Sovereignty under Generative AI: A Theory of Authorship Under Generative Cognitive Delegation

Amir Konigsberg, PhD
Tel Aviv University

Proposed article type: Perspective
Proposed length: approximately 6,500-7,000 words
Proposed display items: 2-3 figures/tables

Synopsis

Human thought has always been supported and shaped by resources beyond the individual. These resources include other minds, shared language, institutions, and tools. Generative artificial intelligence adds something qualitatively different to this cognitive ecology: an interactive system that can participate directly in constructing the explanations, arguments, interpretations, and judgments through which understanding is formed.

This creates a growing problem for the study of human cognition. As human and machine contributions become increasingly interwoven, the quality of an individual's cognitive output can become dissociated from what that individual understands, can reconstruct independently, or contributed to the reasoning underlying it. A person may produce a sophisticated analysis while having performed relatively little of its inferential construction; recognize and improve an argument without having generated its structure; or experience substantial understanding after interacting with an explanation whose organization was primarily supplied by an external system. None of these dissociations is unique to artificial intelligence. What generative AI changes is their architecture. Externally generated reasoning can now participate continuously within the individual's own acts of analysis, writing, explanation, and judgment, making the distribution of cognitive authorship increasingly difficult to perceive.

The proposed Perspective develops *authorship opacity* as the problem created by this condition and *cognitive sovereignty* as the capacity required to address it. Cognitive sovereignty refers to the capacity to maintain meaningful authorship over one's understanding by governing the distribution of epistemically significant cognitive labor between oneself and external systems. The framework identifies three component capacities: provenance calibration, or accurate awareness of how cognitive contributions were distributed; process connection, or meaningful engagement with the reasoning relevant to one's understanding; and delegation control, or the ability to determine what cognitive work to externalize and to reclaim it when circumstances require independent reasoning.

The Perspective we propose will make five connected arguments.

First, it will distinguish human-AI performance from human cognitive competence. Existing evaluations of generative AI frequently ask whether assistance improves the quality, speed, or accuracy of human performance. These are important measures of the capability of the combined human–AI system, but they cannot establish what cognition remains in the human component. As generative systems become more capable, final output quality may become an increasingly poor proxy for human understanding, retained competence, and capacity for independent intervention.

Second, the article will locate this problem within established psychological and philosophical traditions. Cognitive sovereignty overlaps with metacognition, source monitoring, cognitive offloading, critical thinking, epistemic agency, intellectual autonomy, and research on automation reliance, but it asks a distinct question: *how is epistemically significant cognitive labor distributed within a hybrid cognitive system, and does the human participant remain calibrated about and in control of that distribution?* The Perspective will also connect the argument to research on illusions of explanatory depth and externally accessible knowledge, which already demonstrates that subjective estimates of knowledge can diverge from internally available understanding.

Third, the article will address extended cognition as the strongest theoretical challenge to the framework we propose. Rather than treating externally supported cognition as inherently deficient, the Perspective will substantially accept the extended-mind premise that human cognition can legitimately incorporate non-biological resources. It will argue that extended cognition and cognitive sovereignty address different questions: extended cognition concerns the boundaries of the cognitive system; cognitive sovereignty concerns *the organization and governance of cognitive labor within that system*. Cognitive sovereignty can therefore be understood as a property of well-governed extended cognition rather than an argument for cognitive independence.

Fourth, the Perspective will propose a mechanistic model of generative cognitive delegation. Generativity, responsiveness, fluency, personalization, procedural opacity, and extremely low interaction costs allow external systems to move from information retrieval and cognitive scaffolding toward co-generation, inferential delegation, and judgment delegation. The consequences of these arrangements should depend not simply on how much AI is used, but on whether the delegated operations are constitutive of an understanding or competence the person needs to develop or retain. This shifts the relevant independent variable from AI use to the distribution of epistemically significant cognitive labor.

Finally, the Perspective will translate the theory into an empirical research programme. It will propose behavioral measures of provenance calibration, independent reconstruction, transfer, error detection, and adaptive delegation, and develop several predictions. A central experimental paradigm compares *self*-first human-AI interaction, in which individuals construct an initial representation before receiving AI assistance, with *AI*-first interaction, in which a complete generated analysis is supplied before human evaluation. The theoretically critical possibility is that both conditions could produce equally strong final outputs while producing systematically different levels of human understanding, reconstruction, transfer, and authorship calibration.

The Perspective will end by considering the implications for education, expertise, professional judgment, and human-AI interface design. It will argue against a simple opposition between AI use and independent thought. Generative AI can increase cognitive sovereignty when it challenges assumptions, elicits prior judgments, produces counterarguments, or scaffolds human generation. The design objective is therefore neither maximal human effort nor minimal AI involvement, but appropriate human authorship: forms of hybrid cognition in which external capability can expand human performance without making the human participant's cognitive contribution opaque to themselves.

Proposed display items

Figure 1: The Cognitive Sovereignty Model
A conceptual model linking the affordances of generative AI to different distributions of epistemically significant cognitive labor; the three component capacities of cognitive sovereignty; immediate outcomes such as understanding calibration, reconstruction, transfer and error detection; and longer-term outcomes such as retained competence and expertise.

Figure 2: Equivalent Output, Different Cognitive Architecture.
Two human-AI pathways: *self*-first augmentation and *AI*-first generation converge on equally high-quality outputs while producing potentially different levels of human generative participation, reconstruction, transfer and authorship calibration.

Table 1: Cognitive Sovereignty and Adjacent Constructs.
A concise comparison with metacognition, source monitoring, critical thinking, cognitive offloading, epistemic agency, intellectual autonomy and automation trust.

Background, Timeliness, and the Case for Nature

Generative AI is rapidly becoming part of the cognitive infrastructure through which people learn, write, analyse evidence, make professional judgments and formulate decisions. Research has established that these systems can substantially improve human performance in knowledge work, while a growing literature suggests that assisted performance, metacognitive calibration, learning and independent capability do not necessarily move together. At the same time, research on extended cognition, cognitive offloading, source monitoring and externally accessible knowledge provides mature theoretical traditions for understanding how cognition can be distributed beyond the individual.

What is missing is a framework for the new configuration created when the external resource does not merely store, retrieve or transmit information but generates substantial portions of the reasoning itself.

This omission is becoming increasingly consequential. Measures of output quality tell us how well a human-AI system performs, but increasingly less about what the human within that system understands or can do independently. That distinction matters differently but urgently across education, medicine, science, law, organizational decision making and the design of AI systems themselves. As AI capabilities improve, the problem may become harder rather than easier to observe: highly competent systems can maintain excellent external performance while reducing the occasions on which human competence is exercised or tested.

The proposed Perspective brings together bodies of research that are currently developing largely in parallel and identifies a common theoretical problem: the governance of cognitive authorship under generative delegation. It introduces *cognitive sovereignty* not as an argument against distributed or technologically augmented cognition, but as a framework for distinguishing different architectures of human-AI collaboration and generating experimentally testable predictions about their consequences.

The timing is important because current norms and interfaces are being established before the long-term cognitive effects of these architectures are known. Educational institutions are determining when and how students should use generative systems; professions are incorporating AI-generated reasoning into consequential workflows; and AI developers are deciding whether assistants should immediately solve problems or preserve opportunities for human reasoning. A framework capable of distinguishing augmentation from substitution, and governed delegation from authorship opacity, could therefore shape both empirical research and system design.

The topic is appropriate for *Nature* because the problem crosses conventional disciplinary boundaries. It connects cognitive and educational psychology, philosophy of mind and epistemology, human–computer interaction, artificial intelligence, professional expertise and the science of learning. The Perspective's central proposition—that the performance of increasingly capable human–AI systems must be distinguished from the cognition retained by the humans within them, has implications wherever AI becomes part of intellectual work. The objective is to establish a shared conceptual and empirical framework before performance gains obscure the cognitive transformations occurring underneath them.

Selected Recent and Foundational Literature

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*, 7–19.

Clark, A. (2025). Extending minds with generative AI. *Nature Communications, 16*.

Fernandes, D. et al. (2026). AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior, 175*, 108779.

Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General, 144*, 674–687.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*, 3–28.

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature, 627*, 49–58.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*, 676–688.

Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior, 160*, 108386.