

Cognitive Sovereignty: A Theory of Human Authorship in AI-Assisted Cognition

Amir Konigsberg, PhD

Preprint, September 2026

Abstract

Human thought has always depended on resources beyond the individual mind, including other people, language, institutions, representations, and tools. Generative AI introduces a distinctive form of participation in cognition by contributing directly to the explanations, interpretations, arguments, and judgments through which people come to understand, believe, and decide. As these contributions become integrated into human reasoning, it can become increasingly difficult to determine how cognitive work was allocated between person and system, creating what we call opacity in cognitive authorship.

We introduce cognitive sovereignty as a psychological construct describing the capacity to govern the allocation of epistemically significant cognitive work between oneself and external systems while remaining accurate about that allocation, meaningfully connected to the reasoning through which beliefs, judgments, or understanding are formed, and able to adjust or reclaim delegated work when needed. The construct comprises three capacities: cognitive source attribution, reasoning connection, and delegation control.

The framework distinguishes AI-assisted from independent capability, constructing from evaluating reasoning, and felt understanding from reconstructable and transferable understanding. It distinguishes cognitive sovereignty from related constructs, develops testable propositions, and proposes behavioral measures. The central claim is not that extensive AI use is inherently harmful, but that meaningful human authorship depends on

whether the allocation of cognitive work remains visible, understood, and governable.

Keywords: cognitive sovereignty, cognitive authorship, generative AI, cognitive delegation, human-AI cognition

Version note. This manuscript substantially revises and supersedes an earlier preprint posted on SSRN in April 2026 under the title Cognitive Sovereignty: The Authorship Problem in AI-Assisted Thought (SSRN Abstract 6575778). The present version reformulates and substantially extends the theoretical framework, including its component capacities, account of human-AI cognitive allocation, theoretical propositions, operationalization, measurement framework, and proposed empirical tests.

1. Thinking with AI: Who Does What?

Human thought is not a self-contained activity of the individual mind. It unfolds through language and interaction with others and draws on cultural knowledge, institutions, representations, and tools. Much of what any person knows originates beyond them, and intelligent thought often consists not in producing knowledge from first principles but in evaluating, combining, extending, and applying cognitive work that originated elsewhere. A physician deciding on a treatment may combine knowledge acquired through training, evidence produced by researchers they have never met, information contained in a patient's record, and their own clinical judgment.

This distributed character of cognition is one of its strengths. It allows knowledge to accumulate across generations, specialists to develop different forms of expertise, and people to solve problems that exceed the capacities of any single individual. Across education, science, law, medicine, and everyday life, people routinely reason with concepts, evidence, methods, and frameworks they did not themselves produce. The ability to incorporate cognition originating elsewhere is therefore not peripheral to human intelligence; it is one of its defining features.

Generative artificial intelligence can occupy a distinctive role within this distributed architecture by contributing directly to the explanations, interpretations, arguments, and judgments through which people come to believe and understand. Earlier technologies that supported human thought often helped people store, retrieve, represent, transmit, or transform information. Writing allowed thoughts and memories to be recorded outside the mind. Calculators automated formal computation. Search engines located and ranked information in response to a query.

Generative systems can also take on more formative cognitive work. They can identify candidate explanations, compare competing interpretations, formulate hypotheses, synthesize evidence, construct arguments and counterarguments, represent uncertainty, weigh relevant considerations, propose judgments, and articulate conclusions. They can do so interactively, revising their contributions in response to the user's questions, comments, objections, and emerging ideas. The question is therefore no longer simply whether a person used an external resource, but what cognitive role that resource played in producing the resulting understanding or judgment.

Consider the difference between encountering an argument in a book and developing an argument with a generative AI system. Suppose the question is whether a company should adopt a four-day workweek. When a reader encounters an argument in a book, its initial provenance is comparatively clear: another person constructed the argument and the reader encountered it. The reader may criticize, modify, learn from, or eventually adopt it, but they ordinarily retain a reasonably clear sense of where the argument came from.

Interaction with generative AI can be much more iterative and interwoven. A user may begin by asking whether a four-day workweek would improve productivity. The system proposes an analysis; the user challenges an assumption; the system revises the argument. The user introduces evidence from their organization; the system incorporates it, identifies alternative explanations, and reorganizes the trade-offs. Another objection prompts another revision. Over successive exchanges, evidence, interpretations, assumptions, and conclusions are repeatedly reshaped. The final analysis may genuinely reflect both the user's aims and judgments and the system's contributions while making it progressively harder to determine who contributed what to the reasoning.

We use the term *opacity in cognitive authorship* to describe this difficulty in determining how the reasoning was produced and what role the person and the AI system each played in producing it. As generative systems become embedded in the processes through which people construct explanations, arguments, interpretations, and judgments, the allocation of cognitive work underlying an individual's own output may become increasingly difficult for that individual to recognize and manage.

The distinctive issue is not cognitive dependence itself, the endorsement of reasoning one did not generate, or the familiar gap between feeling that one understands and

being able to explain. These are longstanding features of human cognition. What is different in the present setting is the way they can become combined within a single ongoing interaction, and the ease with which externally produced reasoning can become woven into activities the person experiences as their own thinking.

This opacity is consequential because cognitive outputs are routinely treated as evidence about the person associated with them. We infer understanding from explanations, expertise from analyses, competence from performance, and judgment from recommendations and decisions. Responsibility and authority are also often assigned partly on the assumption that the person associated with a judgment understands its basis and can answer for it.

As high-quality outputs become compatible with increasingly different allocations of cognitive work between human and AI, output quality alone becomes a less reliable basis for making these inferences. The problem is therefore both first-person and third-person: individuals may misjudge their own cognitive role, and observers may misjudge the understanding, capability, or expertise of the human associated with an AI-assisted result.

1.1 Assisted and Independent Capability

Human accomplishment has always depended on resources beyond the individual. People rely on other people's expertise, instruments, databases, professional standards, scientific literatures, and computational tools. Generative AI extends this familiar dependence by taking on parts of the reasoning itself.

With AI performing substantial parts of an analysis, a person may produce sophisticated work while having considerably less independent ability to reconstruct the reasoning behind it or apply the same reasoning to a new problem. The result may be genuinely strong, but its quality may reflect the capability of the combined human–AI system more clearly than it reflects the person's own understanding or independent capability.

We call these *assisted capability* and *independent capability*. Assisted capability is what an individual can accomplish as part of a human–AI system. Independent capability is what the individual can explain, reconstruct, diagnose, or perform when the relevant external support is absent, unavailable, or unreliable.

Dario, for example, may show high assisted capability when, working with AI, he produces a sophisticated analysis of whether his company should enter a new market,

incorporating uncertain demand, competitive dynamics, regulatory constraints, costs, and strategic trade-offs. He would show high independent capability if he could reconstruct the reasoning behind the analysis, identify its weaknesses, or apply the same underlying reasoning to a new problem without relying on the AI to perform that work for him.

A difference between assisted and independent capability is not by itself a problem. Tools are valuable precisely because they allow people to achieve more than they could achieve unaided. A person who can frame problems well, direct an AI productively, evaluate its contributions, and integrate them into a strong result may possess an important capability that is visible only in collaboration with the system.

The problem arises when assisted and independent capability are treated as though they were the same thing, or when the person themselves is poorly calibrated about the difference. This matters especially when an AI-assisted output is being used as evidence of what the person understands or can do, as in education, professional assessment, hiring, credentialing, or the delegation of authority.

1.2 Constructing and Evaluating Reasoning

Much of what we know also depends on our ability to evaluate and rely on reasoning we did not work through ourselves. No one independently derives every scientific explanation, historical conclusion, mathematical result, or legal principle they have good reason to accept. Relying on reasoning developed by others is therefore a familiar and often rational feature of human cognition.

Nevertheless, constructing a line of reasoning and evaluating one that has already been developed are different cognitive activities. Constructing an argument requires identifying relevant considerations, deciding which evidence matters and how much weight it should carry, considering possible alternatives, working out relationships among different claims, resolving uncertainty, and determining what conclusions follow. Evaluating an already completed argument begins from a different position because much of this structure has already been supplied.

Generative AI can make it exceptionally easy to move between these roles. A person may begin intending to reason through a problem themselves but gradually shift, through a series of seemingly minor requests for assistance, from constructing the reasoning to evaluating reasoning largely developed by the system. Importantly, the

person may remain cognitively active throughout. Prompting, comparing, criticizing, editing, and choosing are all genuine cognitive acts. Yet this continued activity can make it difficult to notice how much of the work of *constructing* the reasoning has shifted to the AI.

Consider someone deciding between two job offers. Sam identifies the factors that matter to him, explains the trade-offs he sees, and states his initial preferences. He then asks AI to help him think through the choice. Sam has supplied much of the relevant material, but the AI may still decide how those considerations fit together, identify which tensions deserve the most weight, and construct the reasoning through which one option ultimately appears preferable.

Sam may sincerely endorse the final choice and may have contributed extensively to the interaction. Yet contribution, endorsement, and authorship are not equivalent. The important question is not only how much Sam contributed, but what role those contributions played in producing the judgment. A person can provide many of the inputs and accept the conclusion while having played a more limited role in constructing the reasoning that made the conclusion seem warranted.

1.3 Felt Understanding and Reconstructable Understanding

A different question concerns whether feeling that one understands something corresponds to what one can actually reconstruct, explain, or apply. This distinction is especially important with generative AI because one of its clearest benefits is its ability to make complex ideas easier to understand. Explanations can be reformulated, elaborated, and adapted in response to a person's questions and difficulties.

Imagine Sundar asking an AI system to explain why inflation can remain high even after interest rates rise. The system provides a clear explanation of the relevant mechanisms, and the issue now feels intelligible to him. That sense of clarity may reflect a genuine increase in understanding. But the experience of understanding can also exceed what remains available later. If Sundar is subsequently asked, without the explanation in front of him, to reconstruct the causal chain, explain why the effects may be delayed, or apply the same logic to a different economic situation, he may discover that his understanding is less complete than it had seemed.

Research on the illusion of explanatory depth demonstrates that people can believe they understand mechanisms more deeply than they are able to explain them (Rozenblit &

Keil, 2002). The important point here is not that understanding should be equated with verbal explanation. People can understand things they cannot fully articulate, and some forms of understanding are practical, tacit, or expressed more clearly in successful action. Rather, a person's sense of understanding may or may not correspond to the abilities that understanding should support in the relevant context.

Depending on the domain, those abilities may include reconstructing an explanation, applying a principle, recognizing when it fails, detecting an error, or transferring the underlying reasoning to a new problem. Someone may be unable to give a polished verbal account of a technical principle yet still demonstrate understanding by applying it correctly and recognizing its limits. Conversely, if a person feels they understand a principle but cannot reconstruct, apply, or adapt it in any relevant way, their sense of understanding may exceed what they are actually able to use.

Generative systems make this calibration problem particularly important because they make explanation dynamically responsive to the individual. An AI can answer the precise question a user asks, rephrase an explanation when the user is confused, adopt the user's vocabulary, provide analogies, respond to objections, and continue elaborating until the experience of confusion diminishes. This responsiveness can be a powerful aid to learning. But the fact that a user now understands an explanation does not necessarily mean that they have developed an understanding they can sustain without continued AI support.

Taken together, these distinctions point toward a broader problem. Maintaining meaningful authorship over one's thinking is not simply a matter of cognitive skill. It also requires recognizing how cognitive work is distributed, remaining connected to the reasoning through which beliefs and judgments are formed, and being able to manage what work is delegated. We use *cognitive sovereignty* to describe this capacity.

2. Defining Cognitive Sovereignty

2.1 The Concept

We define *cognitive sovereignty* as the capacity to govern the allocation of epistemically significant cognitive work between oneself and external systems, such as AI, while remaining aware of and accurate about that allocation, meaningfully connected to formative reasoning, and able to reclaim or adjust delegated cognitive work when needed.

On this account, cognitive sovereignty concerns meaningful authorship over one's thinking. It does not require independently performing every component of the cognitive work involved. It requires governing the role that externally produced cognition plays in forming one's beliefs, judgments, and understanding.

The approach is deliberately pragmatic. Cognitive sovereignty does not require cognitive purity or self-sufficiency. Human cognition has always depended on other people, inherited knowledge, institutions, representations, and tools. The relevant question is therefore not whether cognition is entirely one's own, but whether a person can meaningfully govern how cognition originating elsewhere enters into their own reasoning.

Ilya, for example, may use AI extensively when evaluating a difficult business decision and nevertheless retain high cognitive sovereignty if he understands which parts of the analysis the AI supplied, remains connected to the reasoning on which the recommendation depends, and can challenge or reclaim that reasoning if the system proves unreliable.

The same is true more generally. A scientist may depend on colleagues, databases, statistical software, instruments, and AI; a physician may rely on imaging technologies they could not build themselves; a student may learn through explanations supplied by a teacher. Reliance on external resources is compatible with cognitive sovereignty when the person understands the relevant contribution of the resource, remains connected to the reasoning that matters for their own judgment or understanding, and can govern how the contribution is used.

Conversely, an individual may appear highly autonomous while displaying weak cognitive sovereignty if they routinely adopt reasoning whose origins, assumptions, structure, and limits they cannot recognize or assess. The central distinction is therefore not between dependence and independence, but between *governed cognitive dependence* and *opaque cognitive dependence*.

2.2 Epistemically Significant Cognitive Work

Not every mental operation is equally important for cognitive sovereignty. What matters is *epistemically significant cognitive work*: cognitive activity that materially shapes what a person comes to understand, believe, interpret, or judge.

Depending on the task, this may include determining how a problem is framed, identifying which evidence is relevant, formulating hypotheses, comparing explanations, synthesizing information across sources, identifying causal relationships, accounting for uncertainty, developing counterarguments, drawing inferences, deciding how much weight different considerations should carry, or forming a judgment or conclusion.

Whether a particular operation is epistemically significant depends partly on what the person is trying to achieve. Arithmetic may be peripheral to a strategic analysis but central to learning mathematics. Summarizing an article may be peripheral when deciding whether it deserves further reading but central when the educational objective is learning how to identify arguments and synthesize evidence.

What counts as epistemically significant cognitive work is therefore task-relative. The relevant question is not simply what the AI did, but whether it took on work that matters for the understanding, capability, judgment, or responsibility at issue.

2.3 Three Capacities of Cognitive Sovereignty

We propose that cognitive sovereignty depends on three capacities: *cognitive source attribution*, *reasoning connection*, and *delegation control*. These concern whether a person can identify where epistemically significant cognitive contributions came from, remain connected to the reasoning underlying the relevant belief, judgment, or understanding, and control what cognitive work is delegated.

Cognitive source attribution: identifying where ideas and reasoning came from

Cognitive source attribution is the capacity to distinguish what one developed through one's own reasoning from what was supplied externally, produced jointly, or received in a more complete form and then endorsed. This includes noticing when important cognitive work has shifted to an external system and distinguishing one's own contribution from ideas, reasoning, or structure supplied or shaped by that system. We refer to the accuracy of these judgments as *provenance calibration*.

Cognitive source attribution requires more than remembering that AI was used. A person may know perfectly well that a generative system contributed to a task while having little sense of how the important cognitive work was actually allocated.

Adrian may, for example, underestimate the system's role in shaping how a situation was interpreted or how a problem was framed. He may remember an argument as his own because he substantially edited it, or regard a conclusion as independently reached because it fit his prior beliefs and he endorsed it immediately. What matters is therefore not simply whether the person knows AI contributed, but how accurately they understand which ideas and reasoning they contributed, which were introduced by the system, and which emerged through interaction.

Cognitive sovereignty does not require precise accounting of every contribution. Human ideas often develop jointly, and it may be impossible to determine exactly where a particular formulation originated or who first introduced every element. What matters is whether people retain a reasonably accurate sense of their own role in the reasoning that emerges, especially when this affects what they can reasonably claim to understand, do independently, or take responsibility for.

Reasoning connection: staying connected to how conclusions are reached

Reasoning connection concerns whether a person remains sufficiently connected to the basis on which a belief, judgment, decision, or understanding is formed.

Suppose Demis must choose between two projects and asks an AI system to analyze the options. The system recommends Project A. Reasoning connection does not require him to reproduce every detail of the analysis. It requires that he understand enough of the basis for the recommendation to know why Project A was favored, what trade-offs were involved, where the important uncertainties lie, and what might justify reaching a different conclusion.

Here, *reasoning* refers to the structure of reasons available to and relied upon by the user: the evidence, assumptions, relationships, alternatives, and considerations that make a conclusion appear warranted. It should not be understood as a transparent account of the internal computational process by which the AI produced its output. The framework does not assume that an AI's explanation reveals its underlying computational process, nor does reasoning connection depend on attributing understanding to the AI. Its concern is the human's relationship to the reasons and considerations that enter into their own beliefs and judgments.

Reasoning connection is also not determined by how active the user appears during an interaction. Someone may retain strong reasoning connection while using AI extensively

—for example, by forming an initial view, asking AI to challenge it, examining relevant evidence, and revising the view in response. Conversely, a person may prompt, edit, compare, and choose repeatedly while leaving much of the work of developing the underlying reasoning to the system.

The distinction is not between activity and passivity. It is between remaining connected to the basis on which an understanding or judgment is formed and merely interacting actively with reasoning largely supplied by AI.

Delegation control: deciding what cognitive work to delegate

Delegation control is the ability to determine which parts of a task to delegate to AI and to revise that arrangement when circumstances change. Knowing what has been delegated is not enough; cognitive sovereignty also requires being able to reclaim or redistribute work when necessary.

Suppose Fei-Fei is deciding whether her company should launch a new product in Europe. She asks an AI system to summarize market data, assess competitors, and identify regulatory risks while she retains responsibility for weighing these considerations together. She initially concludes that the expansion is worth pursuing. She then discovers that the recommendation depends heavily on an assumption that regulatory approval will arrive on schedule, despite new information suggesting a substantial delay. Delegation control is reflected in her capacity to reclaim that part of the analysis, re-examine the regulatory assumptions, and reconsider the decision.

What delegation control requires depends on the task. A person may use AI to retrieve information while retaining interpretation and final judgment. In another setting, they may ask AI to search for weaknesses in an argument they have already developed. And in some cases extensive delegation may be entirely appropriate. Someone who asks AI to turn routine meeting notes into a concise summary may have little reason to preserve every operation involved in doing that task manually.

What matters is therefore not how much or how often a person delegates. The cognitively sovereign person is not necessarily the person who uses AI the least. They are the person who understands what they have delegated, why they have delegated it, and under what circumstances they should change that arrangement.

Taken together, the three capacities provide a framework for describing the human role in AI-assisted cognition. They make it possible to ask more precisely what part the

person played in producing an outcome, what understanding and capability remained with them, and how much control they retained over the cognitive work performed by the system. The framework is not intended to prescribe a single correct allocation between human and AI. Its purpose is to make that allocation visible and therefore available for deliberate choice.

3. Philosophical Foundations

Cognitive sovereignty sits at the intersection of several longstanding questions in psychology and philosophy: how people exercise intellectual self-direction, how they depend on knowledge and reasoning originating elsewhere, how technologies become integrated into cognitive activity, and what it means to exercise agency over the processes through which judgments are formed. No single tradition fully captures the construct developed here. Together, however, several traditions clarify the kind of cognitive agency at issue.

3.1 Kant: Intellectual Self-Direction

From the Kantian tradition, cognitive sovereignty inherits the idea that intellectual maturity involves a willingness and capacity to exercise one's own judgment rather than simply defer to another. In "What Is Enlightenment?", Kant (1784) describes self-incurred tutelage as the failure to use one's understanding without the guidance of another. His *sapere aude*, or "dare to know," captures an ideal of intellectual self-direction.

Importantly, the condition is not simply one of incapacity. Part of the failure lies in relinquishing the exercise of one's own understanding when accepting guidance is easier. This makes the idea especially relevant to generative AI, because cognitive work that would otherwise require sustained attention, uncertainty, and effort can now be handed over quickly and repeatedly.

Oren, for example, may be fully capable of deciding whether to challenge an unexpectedly large repair bill for his car: reading the mechanic's explanation, comparing it with the work that was authorized, checking whether the quoted parts and labor seem reasonable, and deciding whether the charge is justified. Instead, he may photograph the invoice, upload it to an AI system, and ask whether the bill is fair and what he should say. The system may identify questionable charges, formulate the argument, and even draft the message he sends. Oren may understand and approve the result, but

much of the reasoning through which the judgment was formed has been performed for him.

Kant's point should not be interpreted as requiring people to think alone. Intellectual maturity has always depended on education, testimony, scientific institutions, and inherited knowledge. For cognitive sovereignty, the relevant point is more specific: people should retain a meaningful role in forming judgments they ultimately treat as their own.

3.2 Phenomenology: When Assistance Becomes Invisible

Phenomenology contributes a different insight by showing how tools can become experientially transparent when incorporated into skilled activity. Heidegger's distinction between readiness-to-hand (*Zuhandenheit*) and presence-at-hand (*Vorhandenheit*) captures two ways a tool can enter experience (Heidegger, 1927/1962).

In readiness-to-hand, the tool is encountered through its use rather than as an object whose properties are consciously examined. A carpenter using a hammer is ordinarily concerned with driving the nail rather than attending to the hammer's weight, shape, or material. When the hammer works as expected, attention is directed toward the task. When it breaks or becomes unsuitable, it becomes an explicit object of attention.

Merleau-Ponty's phenomenology of embodiment develops a related idea: familiar tools can become incorporated into how a person perceives and acts in the world, functioning less like external objects that must be continuously managed and more like extensions of practical engagement (Merleau-Ponty, 1945/1962).

For cognitive sovereignty, the important point is that successful integration can make a tool less salient precisely because it is working well. This becomes especially important when the tool does not merely support an activity but performs part of the cognitive work within it. If a generative system analyzes evidence, compares alternatives, develops explanations, or constructs arguments, the same transparency that makes interaction easy may also make the division of cognitive labor harder to perceive.

The user's attention remains directed toward the problem they are trying to solve. Meanwhile, the distinction between what they are contributing and what the system is contributing may recede from awareness. Externally produced cognition can therefore become integrated into an activity the person experiences as a continuous process of their own thinking.

Phenomenology thus helps explain how opacity in cognitive authorship can arise. The more seamlessly cognition produced elsewhere becomes integrated into an activity, the less visible the allocation of cognitive work itself may become.

3.3 Testimony and Epistemic Dependence

Epistemology has long recognized that human knowledge depends extensively on testimony (Coady, 1992; Lackey, 2008). A person can be justified in believing that Jupiter is the largest planet in the Solar System without independently reproducing the observations and calculations that support that conclusion. Modern epistemic life would be impossible if every individual had to establish every belief from first principles.

Generative AI therefore does not introduce dependence into an otherwise cognitively self-sufficient world. What changes is the form that dependence can take.

When another person tells us something, we usually know who we are relying on. If Tamar accepts a claim about a medication because Mia, a physician she trusts, tells her, the source of the claim remains clear. Tamar can ask whether Mia is knowledgeable, trustworthy, and in a position to know.

Generative systems can occupy a different position. Their outputs may synthesize material from many sources without making the origins of particular claims or lines of reasoning visible. More importantly, the system may do more than supply information. It may organize evidence, identify what appears relevant, compare alternatives, connect considerations, or construct the reasoning on which the person subsequently relies.

The question is therefore no longer only whether an external source should be trusted. It is also what role externally generated cognition played in shaping the reasoning the person now treats as their own, and whether that role remains sufficiently visible and governable.

3.4 Frankfurt: Reflective Agency

Harry Frankfurt's work on desire, freedom, and agency offers another way of thinking about self-direction. Frankfurt (1971) distinguishes simply having desires from being able to reflect on which desires one wants to govern one's actions. Freedom, on this account, involves not only acting but having a reflective relationship to what guides one's action.

Cognitive sovereignty extends this concern from action to thought. People do not merely arrive at conclusions; they can care about how they arrived at them. They can consider which evidence they took into account, which assumptions they accepted, which parts of the reasoning they worked through themselves or delegated, and whether the resulting judgment arose through a process over which they exercised meaningful control.

Mary, for example, may use AI to assist with a difficult decision while still examining important evidence herself, challenging assumptions she finds unwarranted, and deciding which parts of the system's reasoning should influence her own judgment.

In this sense, cognitive sovereignty concerns *self-determination at the level of thought*. It does not require freedom from external influence. Rather, it requires the capacity to determine the terms on which externally produced cognition becomes part of one's own reasoning and understanding.

Taken together, these traditions point toward a conception of cognitive agency that neither requires independence from external influence nor reduces agency to a final act of endorsement. Agency over thought also depends on the person's relationship to the processes through which their beliefs, judgments, and understanding are formed. Cognitive sovereignty develops this idea for a setting in which generative systems participate directly and interactively in those processes.

4. Cognitive Sovereignty and Related Accounts of Cognition

The value of a new construct depends partly on whether it captures phenomena that are not already adequately described by existing ones. This criterion is especially important for cognitive sovereignty because it overlaps with several established psychological and philosophical constructs. The question is therefore not whether cognitive sovereignty has conceptual neighbors, but whether the particular configuration it describes adds something theoretically and empirically useful.

4.1 Critical Thinking

Critical thinking is commonly understood as involving the ability and disposition to analyze arguments, evaluate evidence, detect fallacies, reason according to appropriate standards, and avoid accepting claims uncritically (Ennis, 1985; Facione, 1990; Halpern, 1998).

Cognitive sovereignty often requires critical thinking, but it asks a different question. Critical thinking asks whether a claim or line of reasoning is well supported, logically coherent, responsive to evidence, and appropriately open to challenge. Cognitive sovereignty asks what role the particular person played in producing that reasoning, which parts were supplied externally, and whether the resulting allocation of cognitive work was understood and governed.

The two therefore need not coincide. A person may be excellent at evaluating an argument generated by AI, spotting weak assumptions and identifying flaws, while having played little role in constructing the argument itself. Conversely, someone may construct an argument entirely on their own and still reason badly. Critical thinking concerns the quality of the reasoning; cognitive sovereignty concerns the person's role in producing, understanding, and governing the reasoning on which they rely.

4.2 Metacognition

Metacognition concerns awareness and knowledge of one's own cognitive processes together with their monitoring and regulation (Flavell, 1979; Nelson & Narens, 1990). Cognitive sovereignty is deeply metacognitive because source attribution and delegation control both require monitoring one's cognitive activity.

AI-assisted cognition, however, introduces an additional problem. When people evaluate how well they have understood or performed, they may partly be evaluating the performance of the combined human–AI system.

In broad terms, metacognition asks, “How well am I thinking, understanding, or performing?” Cognitive sovereignty adds another question: “How much of the thinking I am evaluating was actually mine?” A person may correctly recognize that an AI-assisted result is excellent while still being mistaken about what that result reflects about their own understanding, contribution, or ability to reproduce the reasoning without AI.

Provenance calibration captures one part of this problem: how accurately a person can identify where important cognitive contributions came from as the interaction developed. A user may remember supplying particular facts while incorrectly believing they identified a central risk, comparison, or interpretation that was in fact introduced by the system.

4.3 Source Monitoring

Source monitoring is perhaps the closest established psychological neighbor of cognitive source attribution. The source-monitoring framework developed by Johnson and colleagues concerns how people determine the origins of their memories, knowledge, and other mental experiences (Johnson, Hashtroudi, & Lindsay, 1993).

The overlap is substantial, but cognitive source attribution extends the question from the origin of content to the origin of cognitive work. David may correctly remember that a statistic in his report came from an AI system rather than from a document he read himself. He may nevertheless be much less clear about who recognized that the statistic was important, connected it to other evidence, used it to frame the problem in a particular way, or drew the inference that eventually shaped the conclusion.

The distinction is therefore not only between different sources of information, but between different sources of the reasoning through which information was interpreted and used.

Accurate source monitoring can also coexist with weak reasoning connection. David may correctly report that AI constructed an argument he incorporated into his analysis while being unable to reconstruct why the argument works or develop a comparable argument himself.

Finally, cognitive sovereignty includes a regulatory dimension that source monitoring does not. Knowing where a contribution came from is one thing; being able to decide whether that work should have been delegated and to change the arrangement when needed is another.

4.4 Intellectual Autonomy

Intellectual autonomy concerns the disposition to think for oneself, resist inappropriate conformity, and form beliefs through one's own engagement with reasons and evidence (Siegel, 1988; Baehr, 2011; Roberts & Wood, 2007).

Cognitive sovereignty is concerned with something different. Intellectual autonomy asks whether the judgment a person ultimately accepts is genuinely their own rather than one adopted through deference, conformity, or pressure. Cognitive sovereignty asks how the thinking that produced that judgment was carried out and governed.

The two can therefore diverge. A person may freely reject outside pressure, decide for themselves what they believe, and sincerely endorse a conclusion while still having delegated much of the reasoning that produced it to AI. The system may have framed the problem, identified the relevant considerations, weighed the evidence, and constructed the argument, leaving the person primarily to evaluate and accept the result.

In that case, the individual may remain intellectually autonomous in what they believe while exercising relatively little authorship over how the belief was formed.

Generative assistance also differs from many traditional threats to autonomy because its influence may not feel externally imposed. The contribution is elicited by the user's own question, adapted to their context, and revised in response to their objections. In highly iterative exchanges, it may feel continuous with the person's own thought. Cognitive sovereignty concerns the capacity to recognize and govern this less visible redistribution of cognition.

4.5 Epistemic Agency

Epistemic agency can be understood as the capacity to direct one's pursuit of knowledge through seeking evidence, asking questions, evaluation, belief revision, and responsibility for one's beliefs (Elgin, 2013). A sophisticated AI user may demonstrate substantial epistemic agency by asking strong questions, requesting alternatives, challenging assumptions, and managing a complex investigation.

Yet guiding an inquiry is different from performing the reasoning that produces its conclusions. A person may become highly skilled at directing AI-assisted work while doing less of the inferential work themselves. They may be active, thoughtful, and critical throughout while remaining uncertain about which conclusions they reached through their own reasoning and which they came to accept after the system introduced them.

Cognitive sovereignty therefore asks how cognitive work is allocated within an inquiry, not merely whether the human initiated, directed, or supervised it.

4.6 Cognitive Offloading

Cognitive offloading refers to using external actions or resources to reduce internal cognitive demand (Risko & Gilbert, 2016). Cognitive sovereignty is not opposed to

offloading. Intelligent cognition often depends on deciding what not to do mentally. Writing, calculators, search engines, and other technologies can extend human cognition without threatening forms of understanding a person has reason to preserve.

Not all offloading, however, plays the same role in a task. We refer to offloading as *instrumental* when an external tool performs a supporting operation while the person retains the cognitive work through which the relevant understanding or judgment is formed.

Imagine someone comparing two mortgage offers. They may use a basic calculator to determine the monthly payment and total interest under each option. The calculator performs the computation, but the person still decides whether the lower monthly payment of one mortgage is worth its greater total cost over time, considers how each option fits their financial circumstances, and ultimately decides which mortgage to take. The calculation has been offloaded; the reasoning through which the choice is made remains with the person.

We refer to offloading as *constitutive* when the external system also takes on part of the reasoning through which the relevant understanding or judgment is formed. If a system identifies which differences between the mortgages matter, weighs lower monthly payments against long-term cost, incorporates the person's circumstances, and recommends which mortgage to choose, it is no longer merely performing a supporting calculation. It is participating in the reasoning through which the judgment itself is produced.

The distinction therefore lies not in the technology being used but in the role the tool performs within the task. The same technology may support cognition instrumentally in one context and substitute for a constitutive cognitive operation in another.

4.7 Trust in Automation and Automation Bias

The literature on automation has long examined when people should rely on automated systems, when they rely too much, and when they fail to rely when they should (Parasuraman & Riley, 1997; Lee & See, 2004). Much of this work is concerned with trust and the calibration of reliance: does the person rely on the system to an extent justified by its actual capability and reliability?

Cognitive sovereignty raises a different question. Even when reliance on a system is entirely warranted, what cognitive role does the person retain?

Someone may have good reason to trust a highly reliable system while understanding little of the reasoning involved in the task it performs. Conversely, a person may rely heavily on AI while retaining substantial cognitive sovereignty if they understand what has been delegated, remain connected to the reasoning that matters for their own judgment, and have deliberately determined that the delegated capability need not remain one they personally exercise.

Trust and cognitive sovereignty therefore concern different aspects of the same human–AI relationship. Trust asks whether relying on the system is warranted. Cognitive sovereignty asks whether the person retains self-determination over the role the system plays in their thinking: what they continue to understand and do themselves, what they choose to hand over, and whether they can change that division when needed.

4.8 Illusions of Understanding

Research on the illusion of explanatory depth demonstrates that people can believe they understand mechanisms more deeply than they are able to explain them (Rozenblit & Keil, 2002). Generative AI may extend this problem in a new direction. The issue is not only whether people overestimate how well they understand something, but whether they overestimate how much of the reasoning behind that understanding they themselves produced.

Messeri and Crockett (2024) identify a related risk in scientific practice: AI tools can create conditions in which researchers produce more while understanding less. Cognitive sovereignty focuses on one possible individual-level process through which such a gap may emerge—the miscalibration of cognitive authorship under conditions of generative delegation.

4.9 Existing Conceptions of Cognitive Sovereignty

The phrase *cognitive sovereignty* is not unique to the present account. Recent work uses the term in several related but distinct ways.

Atkinson (2025) uses cognitive sovereignty to describe preservation of a person’s capacity for attention, reflection, and deliberate choice in digital environments designed to predict and shape behavior. His concern is that digital systems may influence not only what people encounter but the temporal conditions under which they respond, reducing opportunities to pause and reconsider.

Klein and Klein (2025) develop the concept within their “extended hollowed mind” framework, which examines the risk that easy access to AI-generated answers may bypass the cognitive effort through which knowledge and reasoning capacities are developed. Their account emphasizes foundational knowledge, metacognitive skill, and willingness to engage in effortful thought as conditions for meaningful intellectual independence.

Branda and Ciccozzi (2026) use cognitive sovereignty in the life sciences, where AI increasingly participates in research, diagnosis, and decision-making. Their emphasis is on retaining the capacity to evaluate evidence, recognize uncertainty, challenge algorithmic outputs, and preserve human responsibility for consequential judgments.

These accounts converge on the importance of preserving human judgment under technological mediation, but they differ in their unit of analysis and the processes they emphasize. The present theory defines cognitive sovereignty more specifically as a psychological capacity for governing the allocation of epistemically significant cognitive work when generative systems participate directly in producing reasoning.

Its contribution therefore lies not in the phrase itself but in an authorship-centered framework consisting of cognitive source attribution, reasoning connection, and delegation control, together with behavioral predictions concerning reconstruction, transfer, provenance, and adaptive delegation.

4.10 Extended Cognition and the Human’s Role Within Hybrid Systems

One influential way of thinking about human–AI cognition comes from work on the extended mind, which asks whether the processes that constitute cognition must be confined to the brain or body.

Clark and Chalmers (1998) argued that, under the right conditions, external resources can become so deeply integrated into cognitive activity that they should be understood not simply as tools assisting cognition but as parts of the cognitive system itself. Clark (2025) has applied this perspective explicitly to generative AI, placing human–AI interaction within the longer history of humans constructing hybrid cognitive systems from biological and non-biological resources.

This perspective is highly relevant to cognitive sovereignty because it shifts attention from whether AI contributes to cognition to how biological and artificial processes become combined within a single cognitive activity. If human–AI cognition can

legitimately be understood as hybrid cognition, however, a further question follows: what is the position of the human thinker within that hybrid system?

From an extended-mind perspective, separating human and AI contributions may initially appear unnecessary. If the extended system is doing the thinking, why insist on identifying which component performed which cognitive operation?

Cognitive sovereignty accepts much of the extended-cognition premise but argues that this further distinction can still matter. Two equally capable hybrid systems can leave their human participants in very different cognitive positions.

Consider two people using AI to decide whether to buy a particular used car. Maya begins by identifying what matters to her: reliability, running costs, safety, and suitability for her commute. She develops an initial view, then asks AI to challenge it, identify risks she may have missed, and provide evidence against her preferred choice. She checks those concerns and revises her judgment where necessary.

Daniel instead gives the AI the car's details, his budget, and his commuting needs and asks whether he should buy it. The system identifies the relevant considerations, compares the risks and benefits, and recommends that he proceed. Daniel reads the analysis, checks that nothing appears obviously wrong, and follows the recommendation.

Both interactions may involve tightly integrated human–AI cognition, and both may produce the same good decision. Yet the human occupies a different functional position within each system. Maya uses AI within a reasoning process she continues to construct and revise. Daniel largely receives and evaluates a reasoning process constructed for him.

These differences matter even when hybrid performance is equivalent. The same output may reflect different levels of human understanding and capacity to intervene. A person who can explain the basis of a judgment, recognize when its assumptions fail, and take over when the system becomes unreliable occupies a different epistemic position from someone who can only evaluate and endorse the result.

Heersmink's (2015) multidimensional account of cognitive integration is useful here. External resources vary in reliability, durability, accessibility, information flow, procedural transparency, and informational transformation. Generative AI can become highly integrated along several of these dimensions while remaining procedurally

opaque. Pritchard (2010) similarly notes that extended cognition complicates the boundary between an agent's contribution and that of an external resource.

Cognitive sovereignty adds another dimension to this discussion: the human agent's capacity to recognize and govern the integration itself.

The extended-mind perspective therefore does not dissolve the problem; it helps specify it. It establishes that thinking with external resources can be cognitively legitimate. It does not establish that every allocation of cognitive work within an extended system places the human in the same cognitive position, nor does it settle what capacities the human should retain when learning, expertise, oversight, responsibility, or independent intervention matter.

Cognitive sovereignty can therefore be understood not as an alternative to extended cognition but as a property of well-governed extended cognition.

Figure 1 summarizes how cognitive sovereignty differs from its closest conceptual neighbors.

Figure 1
Cognitive Sovereignty and Adjacent Constructs

Cognitive Sovereignty and Adjacent Constructs		
Construct	Core question	Cognitive sovereignty asks
Critical thinking	Is the reasoning sound?	What role did I play in producing and governing it?
Metacognition	How well am I thinking or performing?	How much of the thinking I am evaluating was actually mine?
Source monitoring	Where did this information or memory come from?	Where did the cognitive work come from, not just the content?
Intellectual autonomy	Is the judgment genuinely my own?	Did I govern the reasoning that produced it?
Epistemic agency	Am I actively directing the inquiry?	Which parts of the reasoning am I actually doing?
Cognitive offloading	What cognitive work am I externalizing?	Is the delegated work instrumental or constitutive for this task?
Trust in automation	Is reliance on the system warranted?	What cognitive role do I retain when I rely on it?
Illusion of explanatory depth	Do I understand as much as I think I do?	What can I still explain, reconstruct, or apply without AI?
Extended cognition	Can external resources become part of the cognitive system?	What role does the human occupy within that hybrid system?



Cognitive sovereignty adds an authorship question: who did what cognitive work, and can the human govern that division?

© 2026 Amir Konigsberg. All rights reserved.

Note. The figure contrasts the central question asked by each adjacent construct with the additional question introduced by cognitive sovereignty. Cognitive sovereignty focuses on the human role in producing and governing cognition when AI participates in the thinking process.

5. How Cognitive Work Is Allocated Between Humans and AI

If the quality of a human–AI output does not by itself tell us what role the human played in producing it, we need a way to represent how the cognitive work was divided between the person and the system.

AI use itself is too coarse a variable. Two people can use the same system, on the same task, for the same amount of time while playing very different roles in the reasoning process. One may use AI mainly to retrieve information, expose weaknesses, or generate counterarguments to a position they have already developed. Another may rely on it to frame the problem, identify what matters, weigh the evidence, and formulate the conclusion. Both are using AI, but the division of cognitive work is different.

The more theoretically useful variable is therefore how epistemically significant cognitive work is allocated within the hybrid system.

Generative AI is distinctive in two related respects: in the kinds of cognitive work it can perform and in the responsiveness with which that work can become part of a person's reasoning. Its contributions can include generating ideas and explanations, retrieving and synthesizing knowledge, analyzing and comparing information, solving problems, generating alternatives, and proposing next steps or judgments.

Importantly, these contributions are not static. A system can revise an argument in response to an objection, reorganize an explanation around a particular point of confusion, incorporate newly supplied evidence, or generate an alternative interpretation on request. This allows externally generated cognition to participate repeatedly in the development of the person's reasoning, with each contribution shaping what the person considers, challenges, or concludes next.

The low interaction cost of invoking these capabilities also matters. Complex assistance that once required consulting an expert, locating relevant sources, or spending substantial time working through a problem can now be requested repeatedly within a single cognitive episode. A shift in cognitive allocation therefore need not result from a conscious decision to "hand over the reasoning." It may emerge gradually through a series of individually small requests for assistance.

Taken together, these capabilities permit a range of allocations. In *retrieval*, AI supplies relevant information and the person determines how it is interpreted and used. In *scaffolding*, AI helps structure, explain, question, or guide the person's thinking. In *co-generation*, person and AI jointly develop ideas or reasoning. In *reasoning delegation*, the system performs substantial analysis, synthesis, comparison, or inference. In *judgment delegation*, the system also proposes or makes the resulting recommendation or judgment.

A single interaction may move among these forms. They are not ordered stages or levels of quality, nor do they form a normative scale on which greater delegation is necessarily worse. The significance of an allocation depends on the purpose of the activity.

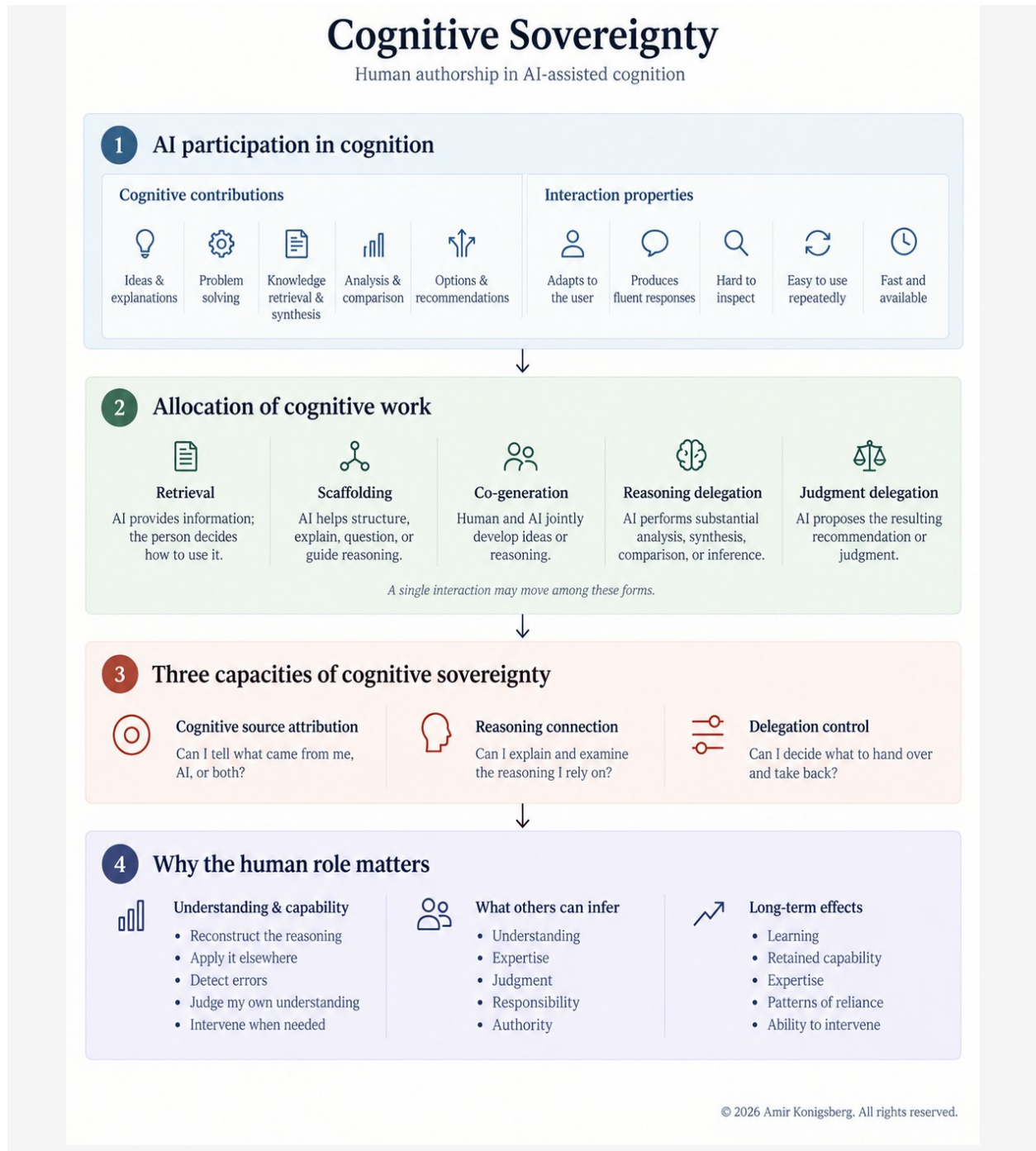
If the objective is to obtain an accurate answer efficiently, extensive reasoning delegation may be entirely appropriate. If the objective is to develop the ability to perform that reasoning independently, the same allocation may bypass the very cognitive work the activity is intended to exercise.

Within this system, cognitive sovereignty concerns how the person relates to the resulting allocation. Cognitive source attribution concerns whether they can distinguish where epistemically significant contributions came from. Reasoning connection concerns whether they remain connected to the reasoning that matters for the relevant understanding or judgment. Delegation control concerns whether the allocation is deliberately managed and can be changed when circumstances require it.

These differences should become visible in outcomes that cannot be inferred from final product quality alone. Two people may produce equally strong analyses but differ in whether they can later reconstruct the reasoning, transfer it to a novel problem, detect a plausible error, identify assumptions on which the conclusion depends, or accurately estimate how much of the thinking they themselves performed.

Figure 2 summarizes the proposed framework, including forms of cognitive allocation, the three capacities of cognitive sovereignty, and the human outcomes they are expected to shape.

Figure 2
Cognitive Sovereignty: Human Authorship in AI-Assisted Cognition



Note. The framework distinguishes what generative AI can contribute, how cognitive work is allocated between human and AI, the three capacities of cognitive sovereignty, and why those differences matter. Allocation modes are not ordered stages; a single interaction may move among them.

6. Empirical Motivation: Evidence for the Proposed Relations

The studies reviewed below provide initial empirical grounding for the framework by showing that several of the patterns cognitive sovereignty seeks to explain are already visible in existing research. Across different settings, AI-assisted performance can

diverge from people's understanding of their own capabilities, reduced cognitive effort can alter engagement and learning, and strong immediate performance can coexist with weaker retained capability.

None of these studies measures cognitive sovereignty as defined here. Their importance lies instead in showing that the relationships brought together by the framework are empirically plausible and sufficiently consequential to warrant direct investigation.

One particularly relevant line of evidence concerns the relationship between assisted performance and people's judgments of their own performance. Fernandes et al. (2025) examined participants solving logical-reasoning problems with AI assistance and found an important divergence: participants performed better with AI, yet their assessments of their own performance were systematically miscalibrated. Greater AI literacy did not eliminate the effect.

The study does not establish that participants confused AI-generated reasoning with their own, because the provenance of particular ideas or inferences was not measured directly. It does, however, show that assisted accomplishment and self-knowledge can diverge. Cognitive sovereignty asks whether one source of that divergence is uncertainty about what the system accomplished, what the person themselves contributed, and what the resulting performance therefore reveals about their own capability.

Earlier research on externally available knowledge points to a related vulnerability. Fisher et al. (2015) found that searching the internet for explanations could inflate people's estimates of how much knowledge they themselves possessed. The finding suggests that the psychological boundary between knowledge available through an external resource and knowledge held internally is already permeable. Generative AI may intensify this ambiguity because the external resource can supply not only information but also the explanatory organization through which that information becomes meaningful.

A second body of evidence concerns the depth of cognitive engagement. Stadler, Bannert, and Sailer (2024) compared students conducting scientific inquiry with LLM assistance and with more conventional information-search tools. Their findings indicate that generative assistance can reduce experienced cognitive demand while changing how deeply students engage with aspects of the reasoning task.

The theoretical implication is not that reducing mental effort is inherently undesirable. Eliminating unnecessary cognitive burden is one of the principal benefits of intelligent tools. The important question is what kind of effort disappears. Effort spent remembering arbitrary details may be expendable; effort involved in comparing evidence, constructing causal relationships, resolving contradictions, or working through uncertainty may be part of the process through which understanding develops. Cognitive sovereignty therefore redirects attention from how much effort AI removes to what cognitive function the removed effort was serving.

Evidence from professional work reveals a complementary pattern. Dell’Acqua et al. (2026) found substantial gains from GPT-4 on consulting tasks within the system’s capability frontier: participants completed more tasks, worked faster, and produced higher-quality solutions. On a task outside that frontier, however, AI-assisted participants were less likely to reach the correct conclusion.

The importance of this result lies not simply in the fact that AI can err. It shows how strongly hybrid performance can depend on the competence of the system. When AI performs well, extensive delegation may be rational and productive. The importance of retained human understanding becomes more visible when the system ceases to be reliable and the person must recognize the failure, question the reasoning, or take over.

Research on learning and retention provides early evidence concerning the longer-term side of the model. Barcaui (2025) found that students given unrestricted ChatGPT access during self-directed learning subsequently performed worse on an unexpected delayed retention test than students using conventional methods.

The finding is compatible with the possibility that immediate access to generated explanations or answers can reduce engagement in some of the processes responsible for durable learning. The mechanism remains unresolved, however. The study did not directly measure how much reasoning students generated themselves, whether they understood the provenance of particular ideas, or whether they strategically regulated assistance. It therefore motivates rather than confirms a cognitive-sovereignty account.

Research on earlier technologies provides an important corrective to overgeneralization. The “Google effect” reported by Sparrow, Liu, and Wegner (2011) is better understood as evidence that people adapt memory strategies to the expected availability of external information than as evidence that search engines straightforwardly degrade memory.

Externalizing cognitive work is not inherently harmful, nor does every form of offloading threaten cognitive sovereignty.

Across these literatures, a common pattern begins to emerge. Generative AI can improve immediate performance while leaving people imperfectly calibrated about their own performance. It can reduce cognitive demand while altering engagement. High-quality outcomes can depend on cognitive work performed by the system rather than retained by the person. And immediate access to capable assistance does not necessarily imply durable independent capability.

The common thread is not that AI simply makes people think better or worse. It is that AI can change the relationship between observable cognitive output and the human cognitive process that produced it. A strong result may tell us increasingly little, on its own, about what the person understood, which parts of the reasoning they performed, or what they could reconstruct or do without the system.

What remains unknown is whether these patterns are linked through the particular architecture proposed here. Cognitive sovereignty predicts that differences in cognitive source attribution, reasoning connection, and delegation control should help explain when assisted performance becomes disconnected from retained understanding, capability, and accurate self-assessment. Testing that prediction requires moving beyond AI use as a binary variable and measuring the allocation of cognitive work itself.

7. Theoretical Propositions

The value of cognitive sovereignty lies not only in providing a language for describing the human role in AI-assisted cognition but also in generating claims that can be tested empirically. The following propositions specify several such implications.

Proposition 1: Performance–capability dissociation. As generative systems perform a greater proportion of the epistemically significant cognitive work required to construct an output, assisted output quality should become a weaker indicator of the human participant’s ability to independently reconstruct, explain, or transfer the underlying reasoning.

The proposition is not that AI assistance necessarily reduces independent capability. Rather, it predicts a weakening relationship between visible performance and what the human can subsequently demonstrate on their own as more of the relevant cognitive work is performed externally.

Proposition 2: Attribution-validity effect. When outputs of comparable quality arise from substantially different allocations of human and AI cognitive work, output quality alone should become a less reliable basis for inferring the human participant's understanding, expertise, or judgment. Making the underlying allocation visible should improve the accuracy of these judgments.

Two individuals may therefore produce work of similar quality while possessing very different levels of independent understanding. Observers who see only the product may treat them as cognitively equivalent even when they are not.

Proposition 3: Construction-evaluation effect. The type of cognitive work the human performs should matter independently of how much activity they display. When relevant information, final output quality, time on task, and overall effort are comparable, people who participate in constructing the reasoning should later show stronger reconstruction, transfer, and error detection than people whose primary role was to evaluate reasoning already constructed by AI.

Constructing a coherent account requires deciding what matters, identifying possible relationships, weighing alternatives, and imposing structure on an initially underdetermined problem. Evaluating an existing account begins after much of that structure has already been supplied.

Proposition 4: Construction pre-emption. The order in which human and AI reasoning occurs should affect the cognitive role the human ultimately plays. When AI supplies a coherent framing or line of reasoning before the person has attempted to construct one themselves, it may reduce the likelihood that the person develops an independent account of the problem at all.

Holding relevant information and final output quality constant, users who first work through a problem themselves and only then receive AI assistance should retain a stronger ability to reconstruct and transfer the reasoning than users who receive a complete AI-generated interpretation at the outset. The critical variable is therefore not only who performs particular cognitive operations, but when those operations occur.

Proposition 5: Provenance-opacity effect. Accuracy in identifying the origins of important cognitive contributions should decline as human and AI contributions become increasingly intertwined through repeated rounds of suggestion, challenge, revision, and incorporation.

This decline should be greater for interpretations, inferences, and ways of framing a problem than for discrete factual inputs. Remembering that a particular statistic came from AI may be relatively straightforward. Identifying who produced an interpretation may be considerably harder if AI first proposed it, the user challenged it, the system revised it, and the user subsequently integrated the revised version into a broader analysis.

The next figure illustrates the logic of Proposition 5 by showing the predicted relation between the degree of human-AI intertwining and errors in authorship attribution.

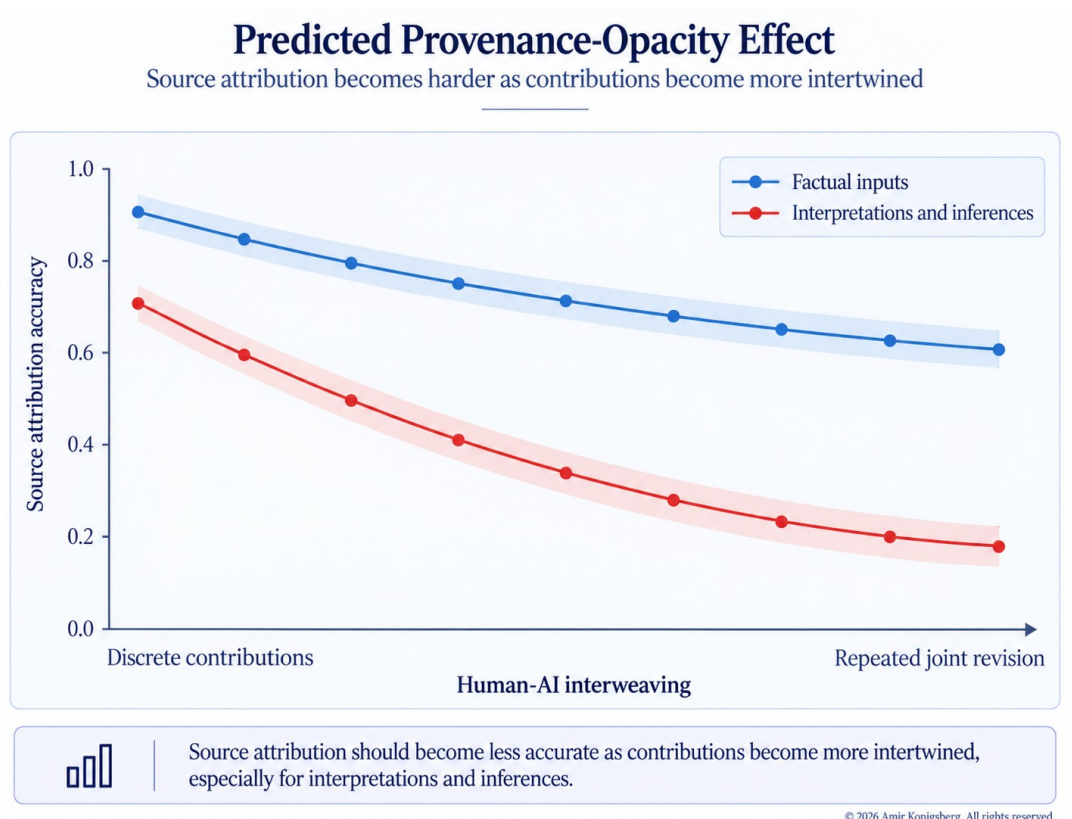


Figure 5: Predicted Provenance-Opacity Effect

Note. As human and AI contributions become more intertwined through iterative suggestion, challenge, revision, and incorporation, provenance accuracy is expected to decline. The decline should be steeper for interpretive contributions, such as framings, inferences, and explanatory structure, than for discrete factual inputs.

Proposition 6: Fluency-induced understanding miscalibration. As AI-generated explanations become more fluent, coherent, and personally tailored, users' subjective sense of understanding should increase more than their later ability to independently

reconstruct or apply the reasoning, particularly when the system has performed much of the work of organizing and connecting the ideas.

Personalization may intensify this effect. An explanation expressed in the user's own terminology, examples, or conceptual language may feel especially familiar and therefore especially easy to experience as part of one's own understanding.

Proposition 7: Expertise as conditional protection. Domain expertise should generally improve the ability to detect AI errors, recognize missing considerations, track important cognitive contributions, and decide which parts of a task can safely be delegated. This protection should reduce rather than eliminate the risk of authorship and understanding miscalibration.

The protection should weaken under conditions that encourage reliance, including high prior system reliability, limited time, and especially fluent or persuasive outputs. Expertise should therefore be treated as a moderator rather than as a guarantee of cognitive sovereignty.

Proposition 8: Adaptive delegation. Higher cognitive sovereignty should be associated not with less AI use but with more responsive control over what is delegated. Individuals with stronger cognitive source attribution, reasoning connection, and delegation control should adjust the type and extent of AI assistance as system reliability, task stakes, learning goals, and their own expertise change.

The relevant capacity is not generalized skepticism or a stable preference for doing things independently. It is the ability to regulate delegation in accordance with the demands of the situation.

Proposition 9: Long-term delegation effect. Repeated delegation of cognitive work that a person still needs to understand or perform independently should, over time, widen the gap between assisted and independent capability, particularly when reasoning connection is weak.

Frequent AI use alone should not produce this effect. AI use that requires the person to construct explanations, generate counterarguments, retrieve relevant knowledge, test assumptions, or apply reasoning in new contexts may maintain or strengthen competence. The predicted long-term effect therefore depends on *what* is repeatedly delegated, not simply on how often AI is used.

Proposition 10: Incremental validity. Behavioral measures of cognitive source attribution, reasoning connection, and delegation control should explain meaningful variation in outcomes such as independent reconstruction, transfer, adaptive reliance, provenance calibration, and AI-error detection beyond that explained by established constructs including metacognitive ability, source monitoring, critical thinking, domain expertise, intellectual autonomy, and trust in automation.

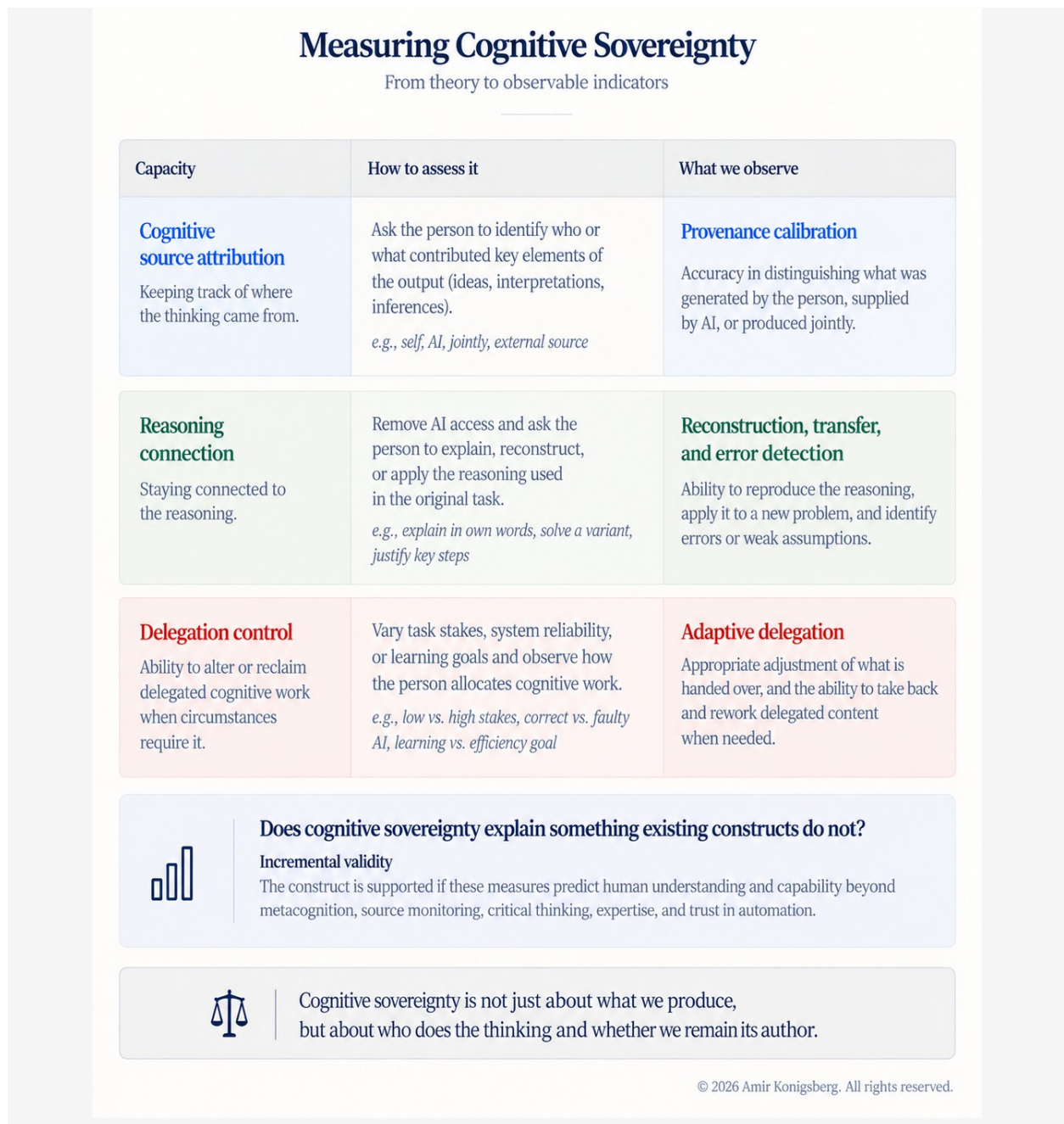
8. Operationalizing Cognitive Sovereignty

The reflexive nature of cognitive sovereignty creates an immediate measurement problem. If part of the construct involves recognizing how much of the thinking a person actually did themselves, then people experiencing the greatest authorship opacity may also be the least able to report that accurately. Self-report therefore cannot by itself establish whether cognitive sovereignty is being maintained or weakened.

A useful measurement strategy should compare what people believe about their cognitive contribution with what their behavior and interaction history indicate.

Figure 3 summarizes this measurement strategy, showing how each of the three proposed capacities can be translated into a behavioral assessment and an observable outcome.

Figure 3
Measuring Cognitive Sovereignty



Note. Cognitive source attribution is assessed through provenance calibration; reasoning connection through reconstruction, transfer, and error detection; and delegation control through adaptive changes in cognitive allocation. The incremental validity of the construct depends on whether these measures explain relevant outcomes beyond neighboring constructs.

One way to measure cognitive source attribution would be to have participants work through a complex reasoning problem with AI while preserving a detailed record of how the interaction developed. The record would show when ideas, interpretations, pieces of evidence, objections, reasoning steps, and conclusions first entered the exchange. After completing the task, participants could be asked to identify where these

contributions came from: which they introduced, which came from the system, and which emerged through interaction.

This would make it possible to distinguish at least two forms of provenance measurement. *Local provenance accuracy* would capture whether participants correctly identify the origins of particular contributions. *Global provenance calibration* would compare their overall estimate of how much of the thinking they performed themselves with the observed allocation across the interaction.

Such a measure would extend beyond conventional source monitoring because it would ask not only where information came from, but who framed the problem, generated a comparison, reconciled conflicting evidence, or constructed the inference through which a conclusion became warranted.

Reasoning connection requires a different kind of measure. After completing an AI-assisted task, participants could lose access to both the conversation and final output and be asked to reconstruct the reasoning in their own words. The aim would not be to test memory for wording. It would be to determine whether they could still explain the argument, identify why particular evidence mattered, recognize important assumptions, and reproduce the relevant relationships without relying on AI.

Reconstruction alone could partly reflect memory for the specific problem. A stronger test would therefore require *transfer*. Participants could receive a new problem that differs on the surface but depends on the same underlying reasoning. Successful application would indicate that something more durable than familiarity with the original output had been acquired.

A further study could examine third-person attribution. Independent observers could evaluate outputs of similar quality while receiving different information about how much of the underlying reasoning was carried out by the person and how much by AI. Observers could then judge the person's understanding, expertise, or quality of judgment. These assessments could be compared with the person's actual performance on independent reconstruction and transfer tasks.

Delegation control requires measuring whether people change their use of AI appropriately when conditions change. A study could begin with a system that performs accurately and consistently, giving participants good reason to rely on it. At a predetermined point, the system could introduce a plausible but important error.

Participants might accept the conclusion, challenge it, seek supporting evidence, reconstruct part of the analysis themselves, request an alternative interpretation, or override the recommendation.

The key measure would be whether the person changes what cognitive work they leave to the system once they have reason to doubt its reliability. Someone who relies appropriately on a reliable system but shifts toward more independent reasoning when reliability deteriorates may demonstrate stronger delegation control than someone who simply distrusts AI from the outset.

A complementary paradigm could vary task difficulty, familiarity, stakes, learning value, and system reliability while allowing participants to choose different levels of AI assistance. The important prediction is that cognitive sovereignty should not appear as uniformly low reliance. A cognitively sovereign participant may delegate more in some conditions and less in others because what matters is whether the allocation fits the goals and circumstances of the task.

Early measurement should therefore resist collapsing cognitive sovereignty immediately into a single questionnaire score. At least initially, provenance calibration, participation in reasoning construction, independent reconstruction, transfer, attribution calibration, and adaptive delegation should remain behaviorally distinct. Their empirical relationships can then determine whether they support the proposed three-capacity architecture or suggest a different structure.

These measures can also be combined within a single experimental design. The following section describes one such initial test.

9. Same Output, Different Human Role: A Critical Initial Test

A straightforward early test of the theory would compare different ways of structuring interaction with AI rather than simply comparing AI use with no AI use.

Imagine three groups asked to solve the same difficult reasoning problem. One group would work without AI. A second, *self-first* group would first be required to write down its own explanation or judgment before consulting the system. After doing so, participants could use AI freely to challenge, extend, reorganize, or improve their reasoning. A third, *AI-first* group would receive a strong AI-generated analysis at the outset and would then evaluate, revise, and improve it.

The most important comparison would be between the two AI-assisted groups. Both could use the same AI, receive the same relevant information, and have the same amount of time. Their final answers might also be similar in quality. The difference would lie in what participants had to do before receiving assistance.

Self-first participants would need to make sense of the problem and develop an initial account themselves. AI-first participants would begin with an already organized explanation and would primarily respond to, evaluate, and modify what the system had provided.

Once final answers were submitted, the AI and completed output would be removed. Participants would then be asked to reconstruct the reasoning, identify why particular pieces of evidence matter, identify important assumptions, apply the same underlying logic to a new problem, and detect an error inserted into a related analysis. They would also report how much of the original reasoning they believed they had contributed, how well they thought they understood it, and how capable they believed they would be of reproducing the analysis without AI.

The key question would be whether the two AI-assisted groups produce similarly strong final answers but differ in what they can later do independently. If the self-first group shows better reconstruction, transfer, error detection, or provenance calibration, one of the paper's central claims would become directly observable: similar visible performance can leave the human participant in very different cognitive positions.

The experiment would therefore test more than whether AI improves performance. It would test whether the same successful result can be produced through different allocations of human and AI cognitive work, and whether those allocations leave different psychological consequences for the human participant.

Figure 4
Same Output, Different Human Role



Note. The figure compares self-first and AI-first ways of structuring interaction. Both may produce similarly strong visible results while leaving the human participant with different levels of independent reconstruction, transfer, error detection, and provenance calibration.

10. Cognitive Sovereignty as a Property of Good Human–AI Collaboration

The framework would be too narrow if it treated generative AI mainly as a threat to human cognition. The same capabilities that allow AI to take over parts of thinking can also be used to deepen and strengthen human thought.

An AI system can ask users to formulate their own position before showing its response, ask them to explain the reasons for a judgment, expose hidden assumptions, generate counterexamples, identify missing evidence, or ask what evidence would cause them to change their mind. Used this way, AI does not have to resolve uncertainty for the user. It can make the structure of the uncertainty more visible and require the user to engage with it.

Generative AI can also function as a cognitive mirror or adversarial collaborator. It can help a person articulate partially formed intuitions so they can examine them, point out contradictions across a long discussion, raise objections the person had not considered, and push them to defend or revise their position.

These uses make clear that cognitive sovereignty is not preserved simply by keeping AI at a distance. In some circumstances, deeper interaction may strengthen it. A person who asks AI to challenge their reasoning, expose weaknesses, and generate alternatives may become more connected to the basis of their own judgment than someone who reasons entirely alone but never seriously tests it.

The relevant distinction is therefore not between AI-assisted and unaided thought. It is whether AI replaces cognitive work the person has reason to perform or retain, or instead helps the person engage more deeply with that work.

This also suggests a more useful design objective than simply keeping a “human in the loop.” A human may remain formally involved while contributing very little to framing the problem, weighing evidence, or constructing the reasoning. The more important question is what cognitive role the human actually plays within the loop.

The goal is not to maximize human effort or require people to perform every operation themselves. It is to preserve appropriate human authorship over the forms of thinking that matter.

11. Boundary Conditions

The effects of delegating cognitive work to AI should depend on several important boundary conditions.

First, expertise matters. Experts possess richer representations of their domains, stronger expectations about what good reasoning should look like, and larger bodies of knowledge against which AI-generated claims can be checked. These resources should

make it easier to detect errors, notice missing variables, or recognize inappropriate framings.

Expertise may nonetheless create a paradox. As AI systems become more capable, experts may need to exercise their own expertise less often. A radiologist, lawyer, programmer, or analyst may remain capable of performing a task independently while increasingly choosing not to do so because working with AI is faster and produces results that are at least as good.

Expertise therefore serves both as protection against inappropriate reliance and as a capacity that may be exercised less frequently precisely because AI assistance is effective. Studies measuring performance during AI-assisted tasks can show whether experts become more productive, but they cannot by themselves show what happens to expertise after years in which important components of the relevant reasoning are rarely exercised.

Second, the purpose of the task matters. The same delegation can be useful in one setting and counterproductive in another. An expert who asks AI to summarize dozens of papers before deciding which ones deserve closer reading may be using their attention wisely. A student whose assignment is intended to teach them how to compare and synthesize those papers may bypass the very cognitive work the exercise is intended to develop by making the same request.

Third, stakes and responsibility matter. Many everyday decisions provide little reason to preserve independent competence. Someone can follow a friend's restaurant recommendation without understanding the friend's reasoning. The situation is different when responsibility for the outcome still rests with the person.

A doctor relying on an AI-generated treatment recommendation, a scientist using AI to interpret evidence, a developer approving AI-generated code for an important system, or a professional making a consequential judgment may need sufficient reasoning connection to recognize when the system's analysis is off the mark. The greater the responsibility retained by the human, the more important it becomes that they preserve enough understanding to identify problems and intervene.

Fourth, system reliability matters. Greater reliance on an AI that consistently performs well can be entirely rational. Cognitive sovereignty does not require resisting reliable

assistance. The difficulty is that highly reliable systems may give users fewer opportunities to discover whether their own capability has weakened through disuse.

The real test may arise when the system encounters unusual conditions, a problem outside its competence, or changing information. What matters is whether the user can recognize when reliance is no longer justified and shift the allocation of cognitive work accordingly.

Interface design can also shape these conditions. Showing a polished answer immediately may prevent the user from forming an initial view; asking for an initial judgment before revealing the system's recommendation preserves that opportunity. Presenting a single recommendation may narrow the task to evaluation, whereas presenting competing possibilities may require comparison.

This points to a role for small amounts of purposeful friction. In conventional interface design, additional steps are often treated as inefficiencies to eliminate. In cognitive systems, a carefully placed pause may preserve an opportunity for thinking that instant assistance would otherwise remove. This idea is compatible with the broader literature on desirable difficulties in learning (Bjork & Bjork, 2011), but the principle here is narrower: friction is useful when it preserves cognitive work the person has reason to exercise or retain.

Finally, individual cognition takes place within institutional environments. Schools, firms, hospitals, and scientific organizations shape how cognitive work is divided through workflow design, default settings, performance incentives, and decisions about which systems people are expected to use. Cognitive sovereignty may therefore depend partly on whether institutions preserve meaningful opportunities for people to exercise, reclaim, and maintain the capabilities that matter.

12. The Developmental Problem

The effects of handing cognitive work to AI may be especially consequential during childhood and adolescence.

Most adults encountered generative AI only after developing foundational abilities such as writing, explaining, building arguments, synthesizing information, and solving problems through years of practice. AI may now change how frequently those abilities are exercised, but in many cases the underlying capacities were already established.

Children growing up with generative AI inhabit a different cognitive environment. AI-generated explanations, arguments, and solutions can be available before the corresponding human abilities have fully developed. This creates an important distinction between losing a capability and never fully developing it in the first place.

An established capacity can weaken when it is rarely exercised. A developing capacity, however, may never become fully established if AI repeatedly performs the very cognitive work through which that capacity would ordinarily be learned.

This does not mean that young people should receive less help or that assistance itself is harmful. Good teaching has always involved support. Teachers simplify difficult problems, demonstrate ways of reasoning, provide partial structures, ask guiding questions, and gradually ask students to take on more of the work themselves.

Generative AI could make this kind of support more available, responsive, and individually tailored than ever before. The developmental question is therefore not whether assistance should be provided, but which parts of the thinking a young person needs to perform themselves, and at what stage, for the underlying ability to develop.

The same AI system may therefore substitute for cognition in one interaction and scaffold its development in another. If AI writes an essay for a student, it may bypass substantial parts of the thinking involved in learning to construct an argument. If it instead asks the student to formulate a claim, give reasons, respond to objections, and revise the argument, it may help strengthen those same capacities.

Research on developmental cognitive sovereignty should examine how well young people can distinguish recognizing that an answer makes sense from having worked it out themselves; whether early reliance affects later transfer to new problems; whether some developmental stages are especially sensitive to construction pre-emption; and whether AI systems can explicitly teach users to recognize and regulate their own cognitive delegation.

Cognitive sovereignty need not therefore be understood only as something technology may weaken. It may itself become a capability that well-designed technology can help people develop.

13. Falsification, Construct Validity, and Limitations

Periods of rapid technological change encourage the proliferation of new terminology. A compelling concept can create an appearance of novelty even when the underlying

phenomenon is already well captured by established constructs. Cognitive sovereignty should therefore be tested rigorously against the possibility that it is simply a new label for things psychology already understands.

The most important test is conceptual and incremental validity. Cognitive source attribution may turn out to be little more than source monitoring applied to human–AI interaction. Delegation control may be explained by existing accounts of metacognitive control and automation reliance. Reasoning connection may overlap so strongly with active learning, depth of processing, or familiar benefits of constructing an answer oneself that it contributes little explanatory value.

If behavioral measures of cognitive source attribution, reasoning connection, and delegation control do not explain meaningful differences in reconstruction, transfer, calibration, error detection, or adaptive reliance after these neighboring constructs are taken into account, there would be little basis for treating cognitive sovereignty as a distinct psychological capacity.

The framework also makes more specific claims that can fail. If people who primarily evaluate high-quality AI-generated reasoning later reconstruct it, transfer it to new problems, and diagnose its weaknesses just as well as people who helped construct the reasoning themselves, then the proposed construction–evaluation distinction matters less than predicted.

Similarly, authorship opacity may turn out to be less significant than proposed. People may remain quite good at identifying the origins of important ideas, interpretations, and inferences even during long and highly interactive exchanges. If so, one of the central phenomena motivating cognitive source attribution would be weakened.

The third-person attribution claim is also testable. If differences in the human contribution do not systematically make judgments of expertise or understanding less accurate, or if revealing the underlying allocation does not improve those judgments, then the proposed social consequence of authorship opacity should be revised.

Delegation control must likewise predict actual adaptation rather than generalized attitudes toward technology. A measure of delegation control should help explain whether people change what they hand over when AI becomes less reliable, the stakes change, or the purpose of the task shifts. If it merely identifies people who distrust AI or prefer doing things themselves, it would not be capturing the construct proposed here.

Several additional limitations and unresolved questions remain.

First, the three-part architecture is conceptual. Cognitive source attribution, reasoning connection, and delegation control may not emerge empirically as separable factors. One may prove causally prior to another, or they may be better understood as interacting processes rather than components of a single capacity.

The breadth and stability of cognitive sovereignty are also unknown. A person may govern AI-assisted cognition well in one domain in which they possess substantial expertise and poorly in another. The construct may vary strongly across situations, or some aspects may behave more like stable individual differences.

Second, instrumental and constitutive delegation are task-relative. No cognitive operation is constitutive in itself. Summarizing a document may be a routine supporting operation for one person and the very capability another is trying to acquire. Empirical work must therefore specify the person's objective before determining whether a particular form of delegation matters.

Third, authorship is distributed even without generative AI. Language supplies conceptual categories; education shapes the forms of argument people regard as legitimate; intellectual communities influence which problems receive attention; and collaboration routinely makes the origins of ideas difficult to disentangle. Cognitive sovereignty must therefore avoid relying on an idealized picture of a person who generates their thinking entirely on their own.

Its claim is narrower: the fact that cognition has always been socially distributed does not mean that every way of distributing cognitive work is psychologically equivalent, or that the human's position within that distribution is unimportant.

Fourth, cultural differences may matter. Cultures and communities differ in how strongly they value individual intellectual authorship, reliance on expertise, shared knowledge, and independent judgment. Cross-cultural research will be needed to distinguish general psychological processes from culturally specific expectations concerning how cognition ought to be organized.

Fifth, the relationship between ability and delegation is likely to run in both directions. People with lower ability may be more likely to delegate work to AI, while repeated delegation may also reduce opportunities to practice and develop the ability.

Longitudinal studies and careful measures of initial competence will therefore be needed to establish causal direction.

Sixth, human and AI contributions may sometimes become genuinely joint. Suppose an AI asks a question that leads a user to an insight they would not otherwise have reached. The idea may have been articulated by the person even though the system played an important causal role in bringing it about. Conversely, the system may reshape a user's initial intuition so extensively that the final argument is genuinely co-produced.

Measures of cognitive authorship will therefore need to represent joint emergence rather than forcing every contribution into a binary human-or-AI category.

Finally, the framework should remain neutral on whether advanced AI systems themselves genuinely understand or reason. Cognitive sovereignty does not depend on denying machine cognition. If future systems become increasingly agentic, the question becomes more explicit rather than less: what role does the human play within the resulting cognitive system, do they understand that role, and can they still govern it?

These are not peripheral caveats. They are tests the framework must pass. Cognitive sovereignty is theoretically valuable only if it helps explain and predict something about human-AI cognition that existing concepts do not already explain.

14. Conclusion: The Governance of Distributed Thought

Human cognition derives much of its power from its capacity to extend beyond the individual. Language, testimony, institutions, representations, and tools allow people to inherit knowledge they did not produce, coordinate forms of expertise they do not personally possess, and solve problems that exceed the capacities of an isolated thinker.

Generative AI expands this architecture by introducing systems that can participate directly in the production of explanations, interpretations, arguments, and judgments.

This capability can increase human performance dramatically. It may democratize access to expertise, accelerate intellectual work, improve expression, expose alternatives that would otherwise remain unseen, and support forms of cognitive collaboration that are genuinely superior to unaided thought.

It also weakens a relationship that has often been taken for granted: the quality of a cognitive product can increasingly tell us less about the cognitive participation of the human associated with it.

A sophisticated argument may demonstrate the capability of a human–AI system without revealing whether the human could reconstruct the reasoning behind it. A compelling explanation may produce durable understanding or merely a powerful experience of clarity. A defensible judgment may reflect deep human reasoning, sophisticated evaluation of externally generated reasoning, or some combination of the two.

The problem is not that human cognition has ceased to be independent. It never was. The problem is that the allocation of cognitive work can become increasingly opaque to the person whose understanding, judgment, capability, or responsibility is at stake.

This paper has proposed cognitive sovereignty as the capacity to govern that allocation. Cognitive source attribution concerns whether people remain accurately oriented toward where epistemically significant cognitive contributions originated and how they were transformed. Reasoning connection concerns whether they remain meaningfully connected to the reasoning necessary for the understanding or capability they seek to retain. Delegation control concerns whether they can deliberately determine what cognitive work to delegate and change that allocation when objectives, stakes, or reliability conditions change.

None of these capacities requires cognitive isolation. A person can rely extensively on AI while remaining cognitively sovereign. Nor does the framework privilege unaided thought as inherently more valuable. In many contexts, refusing capable external cognition would be inefficient or irrational.

The relevant distinction is between distributed cognition that remains visible and governable to the human participant and distributed cognition whose internal organization becomes progressively opaque.

This reframing changes what psychology should attempt to measure. The central question is no longer simply whether AI makes people perform better or worse. It is whether different ways of achieving AI-assisted performance leave people with different levels of understanding, retained capability, knowledge of their own contribution, and capacity for independent intervention.

The challenge ahead is therefore not to preserve an artificial boundary between human and machine cognition. It is to understand how cognition is organized across that boundary and what different arrangements mean for the human within it.

The future of human thought is likely to be increasingly distributed. The psychological question is whether humans remain able to see and govern their place within it. That is the problem of cognitive sovereignty.

References

- Atkinson, R. (2025). Cognitive sovereignty and neurocomputational harm in predictive digital platforms. *Ethics and Information Technology*, 27, 66.
- Baehr, J. (2011). *The Inquiring Mind: On Intellectual Virtues and Virtue Epistemology*. Oxford University Press.
- Barcaui, A. (2025). ChatGPT as a cognitive crutch: Evidence from a randomized controlled trial on knowledge retention. *Social Sciences & Humanities Open*, 12, 102287.
- Branda, F., & Ciccozzi, M. (2026). The comfort of automation: Why cognitive sovereignty matters in AI-driven life sciences. *Artificial Intelligence in the Life Sciences*, 9, 100158.
- Clark, A. (2025). Extending minds with generative AI. *Nature Communications*, 16, 4627.
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19.
- Coady, C. A. J. (1992). *Testimony: A Philosophical Study*. Oxford University Press.
- Dell’Acqua, F., McFowland, E., III, Mollick, E. R., Lifshitz, H., Kellogg, K. C., Rajendran, S., Krayner, L., Candelon, F., & Lakhani, K. R. (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science*, 37(2), 403–423.
- Elgin, C. Z. (2013). Epistemic agency. *Theory and Research in Education*, 11, 135–152.
- Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership*, 43(2), 44–48.
- Facione, P. A. (1990). *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction*. California Academic Press.
- Fernandes, D., Villa, S., Nicholls, S., Haavisto, O., Buschek, D., Schmidt, A., Kosch, T., Shen, C., & Welsch, R. (2025). 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(3), 674–687.
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906–911.
- Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy*, 68(1), 5–20.
- Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist*, 53(4), 449–455.

- Heersmink, R. (2015). Dimensions of integration in embedded and extended cognitive systems. *Phenomenology and the Cognitive Sciences*, 14(3), 577–598.
- Heidegger, M. (1927/1962). *Being and Time* (J. Macquarrie & E. Robinson, Trans.). Harper & Row.
- Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, 114(1), 3–28.
- Kant, I. (1784). An answer to the question: What is enlightenment? *Berlinische Monatsschrift*.
- Klein, C. R., & Klein, R. (2025). The extended hollowed mind: Why foundational knowledge is indispensable in the age of AI. *Frontiers in Artificial Intelligence*, 8, 1719019.
- Lackey, J. (2008). *Learning from Words: Testimony as a Source of Knowledge*. Oxford University Press.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80.
- Merleau-Ponty, M. (1945/1962). *Phenomenology of Perception* (C. Smith, Trans.). Routledge.
- Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627, 49–58.
- Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 26, pp. 125–173). Academic Press.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253.
- Pritchard, D. (2010). Cognitive ability and the extended cognition thesis. *Synthese*, 175(Suppl. 1), 133–151.
- Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688.
- Roberts, R. C., & Wood, W. J. (2007). *Intellectual Virtues: An Essay in Regulative Epistemology*. Oxford University Press.
- Rozenblit, L., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26(5), 521–562.
- Siegel, H. (1988). *Educating Reason: Rationality, Critical Thinking, and Education*. Routledge.
- Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043), 776–778.
- 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.

