Cognitive Sovereignty: The Authorship Problem in AI-Assisted Thought

Amir Konigsberg

Abstract

The rapid integration of large language models into everyday cognitive tasks has created a need for conceptual frameworks adequate to the cognitive consequences of delegating thinking to AI systems. Existing constructs in psychology and also in epistemology, including critical thinking, metacognition, intellectual autonomy, and epistemic agency, each address related phenomena but none adequately captures the specific capacity threatened by habitual AI-assisted cognition, which I define as the capacity to maintain genuine authorship over the process by which one arrives at one's beliefs, judgments, and conclusions. This paper introduces cognitive sovereignty as a distinct construct, defined as the capacity to (a) notice when one's thinking is being displaced, (b) maintain a meaningful connection to how one's beliefs and judgments are formed, and (c) distinguish between genuine reasoning and the subjective impression of having reasoned. I trace the concept's philosophical lineage, engage with the extended mind objection, differentiate cognitive sovereignty from adjacent constructs through systematic comparison, and present a growing body of empirical evidence that motivates the construct. The paper argues that cognitive sovereignty names a phenomenon that existing constructs individually fail to capture and that its articulation is a prerequisite for empirical research on AI's impact on human thinking.

Keywords: cognitive sovereignty, AI-assisted cognition, metacognition, cognitive offloading, epistemic agency, authorship, large language models

1 1. Introduction

Over one billion people now regularly interact with conversational AI systems to assist with tasks that were, until recently, understood as constitutive of human thinking: providing explanations, building arguments, analyzing evidence, forming judgments, making sense of complex situations, and articulating what they believe and why (Chatterji et al., 2025). The technology available to do this is remarkably capable. Large language models produce text that is fluent, contextually appropriate, and

often substantively useful. Users report increased productivity, improved clarity of expression, ready access to information and expertise that would otherwise require extensive research or consultation, and the subjective experience of being understood by a system that appears to grasp what they mean, what they need, and who they are (De Freitas et al., 2025).

These gains are real, and this paper does not dispute them. What this paper does argue is that the habitual delegation of cognitive work to AI systems carries costs that are difficult to detect in the moment, that accumulate over repeated use, and that existing psychological and philosophical constructs are not well positioned to describe. A student who routinely uses the likes of Claude, ChatGPT, or Gemini to write essays may retain critical thinking skills in principle but may gradually lose the capacity to notice when an argument has been assembled for them rather than by them. A physician who relies on AI-generated differential diagnoses may maintain clinical knowledge but may find that her independent diagnostic judgment, the relatively slow, effortful process of working from symptoms to conclusions, has atrophied. A lawyer who uses AI to draft legal briefs may still recognize good legal reasoning when he reads it, but may struggle to produce it without assistance.

What is at risk in each of these cases is not a single identifiable skill but a compound capacity, composed of three interrelated abilities: the ability to notice when one's thinking is being performed by an external system rather than supported by it, to maintain an active awareness of the reasoning process through which one's beliefs and judgments are formed, and to be able to tell the difference between a conclusion one has arrived at through one's own cognitive effort and one that has been absorbed from a fluent, ready-made source. I call this capacity cognitive sovereignty.

The term is introduced not as a synonym for critical thinking, metacognition, intellectual autonomy, or epistemic agency, though it draws on all of these, but as a construct that names something none of these individually capture, which is the specific vulnerability that emerges when the medium in which humans think, language, is shared with a technology able to operate in that medium without any of the experiential grounding that gives language its meaning for the humans who use it, and when that technology's outputs are fluent enough to be mistaken for the products of thought.

This paper has three aims. First, I define cognitive sovereignty and trace its philosophical and psychological antecedents, engaging with the extended mind thesis as the strongest objection to the construct. Second, I systematically distinguish cognitive sovereignty from adjacent constructs, showing where it overlaps and where it diverges. Third, I review a growing body of empirical evidence that motivates the construct and outline the research questions it enables. The goal is to establish cognitive sovereignty as a conceptually coherent and empirically motivated construct, so

that claims about what AI does to human thinking can be investigated rather than merely asserted.

2 2. Defining Cognitive Sovereignty

2.1 2.1 The Concept

Cognitive sovereignty is the capacity to maintain genuine authorship over the process by which one arrives at one's beliefs, judgments, and conclusions. More precisely, it comprises three interrelated abilities:

Displacement detection: the ability to notice when a tool or system is constructing one's reasoning, forming one's judgments, or producing one's conclusions, rather than providing information, prompts, or alternatives that the person then works through independently.

Process connection: the maintenance of an active, first-person engagement with how one's beliefs and judgments are formed, including awareness of the reasoning, evidence, and uncertainty involved in arriving at a conclusion, as well as the cognitive effort the process required.

Authorship discrimination: the ability to distinguish between a conclusion one has reached through one's own reasoning, by weighing evidence, working through uncertainty, and forming a judgment, and a conclusion one has absorbed from an external source and retrospectively endorsed, because the output is coherent, well-structured, or consistent with one's existing views.

The construct is situated at the intersection of metacognition (knowing about one's own cognitive processes), epistemic responsibility (the obligation to form beliefs through appropriate means), and what might be called cognitive self-determination (the ongoing commitment to authoring one's own understanding rather than outsourcing it). The connection to self-determination theory (Ryan & Deci, 2000, 2017) is worth making explicit: SDT identifies autonomy as a basic psychological need and characterizes it as the need to experience one's behavior as volitional and self-endorsed. Cognitive sovereignty can be understood as the epistemic expression of this need, the need to experience one's understanding as genuinely self-authored. If SDT is correct that autonomy is essential to motivation and psychological functioning, then the erosion of cognitive sovereignty may carry costs that extend beyond the epistemic into the motivational and affective domains.

What makes cognitive sovereignty a distinct construct, rather than a combination of existing ones, is its specificity to the conditions created by conversational AI: a technology that responds to users in natural language, that adapts to their context and apparent needs, and whose responses are sufficiently fluent and relevant that they ac-

tivate the same trust responses that evolved for human conversation (Dennett, 1987; Tomasello, 2008; Epley et al., 2007).

These three components are not proposed as empirical dimensions awaiting factor analysis. They are conceptual distinctions that identify the different capacities the construct comprises. A person may have strong displacement detection (they notice when AI is doing the work) but weak process connection (they no longer experience the effortful process of reasoning through problems themselves). A person may have strong process connection in domains where they work without AI but weak authorship discrimination when reviewing AI-assisted output: after reading a legal brief they produced with AI assistance, for instance, they may be unable to identify which arguments they constructed and which the system generated and they subsequently approved. The components are separable in principle, which is what gives the construct internal structure, but the relationship between them, whether they are hierarchical, compensatory, or independently variable, is an empirical question this paper motivates but does not resolve.

2.2 2.2 Philosophical Antecedents

The concept of cognitive sovereignty draws on several philosophical traditions, though it is not reducible to any of them.

From the Kantian tradition, it inherits the idea that intellectual maturity consists in thinking for oneself (*sapere aude*). Kant's (1784) essay *What Is Enlightenment?* describes the condition of self-incurred tutelage, the failure to use one's own understanding without the guidance of another, and attributes this failure not to lack of understanding but to lack of resolve, by which he means the willingness to endure the effort and discomfort of thinking independently when a more convenient alternative is available. The parallel to habitual AI reliance follows naturally from Kant's observation. The availability of fluent conversational AI assistance makes it increasingly easy to forgo the effort of thinking independently, even for those fully capable of doing so.

From the phenomenological tradition, particularly Heidegger (1927/1962) and Merleau-Ponty (1945/1962), cognitive sovereignty inherits attention to the difference between understanding-through-doing and understanding-through-receiving. Heidegger's distinction between *Zuhandenheit* (readiness-to-hand) and *Vorhandenheit* (presence-at-hand) is relevant here: a tool that works seamlessly, such as a well-made hammer or a fluent word processor, ceases to be an object of attention and becomes transparent to the user, who attends to the task rather than the tool. This transparency is generally productive. But for tools that perform cognitive work, the same transparency becomes a liability, because when the tool disappears from awareness, so does the boundary between one's own thinking and the tool's contribution. The user experiences the result

as a single, continuous cognitive process, when in fact a significant portion of it was performed externally.

From epistemology, the concept draws on debates about epistemic dependence and testimony. If I accept a claim on the basis of another person's testimony, I may be justified in my belief, but my justification depends on the epistemic credentials of the source I am relying on (Lackey, 2008; Coady, 1992). When the source is an AI system that has no epistemic credentials in the relevant sense, in that it does not understand what it is saying and has no relationship to truth beyond statistical correlation, the epistemological status of beliefs formed through reliance on that system becomes deeply unclear. Cognitive sovereignty names the capacity to remain aware of this dependency and its implications.

From Frankfurt's (1971) work on freedom of the will, cognitive sovereignty borrows the idea that what distinguishes autonomous agents is their capacity for second-order reflection, which means they have desires about their desires, and can evaluate whether to endorse or resist the desires they find themselves with. A person who craves a cigarette has a first-order desire; a person who craves a cigarette but wishes they didn't, and acts on that wish rather than the craving, is exercising second-order reflection, stepping back from their own mental states and deciding which ones to identify with. Applied to cognition, the analogue extends beyond thinking itself to having a relationship to one's thinking: wanting to be the kind of thinker who arrives at conclusions through genuine reasoning rather than passive absorption, and being able to notice when one has drifted from the first to the second. Cognitive sovereignty, in this sense, is a second-order cognitive capacity, the capacity to monitor and maintain one's relationship to one's own first-order thinking.

2.3 2.3 The Extended Mind Objection

The extended mind thesis (Clark & Chalmers, 1998) provides the strongest philosophical objection to cognitive sovereignty as a construct and therefore requires sustained engagement. If cognitive processes can legitimately extend beyond the boundaries of the individual brain, or beyond the skull, as Clark and Chalmers put it, into notebooks, smartphones, and other external tools, then reliance on AI may simply represent a further extension of cognition rather than a threat to it. On this view, the person who thinks with AI is not losing cognitive sovereignty any more than the person who thinks with a notebook is. The tool is part of the cognitive system, and the outputs of the extended system are genuinely the person's thoughts.

I take this objection seriously and respond to it on several levels. First, Clark and Chalmers's original parity principle holds that if an external process functions in a way that we would count as cognitive if it occurred in the head, then it is cognitive

regardless of where it occurs. The key qualification is functional equivalence. A notebook that stores beliefs the way biological memory does can plausibly be part of an extended cognitive system. But an LLM that generates novel arguments, synthesizes evidence, and constructs conclusions does not function the way any internal cognitive process does. It does not store the person's existing beliefs and retrieve them on demand; it produces arguments, explanations, and judgments that the person did not arrive at through their own reasoning and may never have formed independently. The parity principle does not straightforwardly apply to a tool that produces reasoning and conclusions on the user's behalf rather than storing or extending the user's own cognitive capacities.

Second, Heersmink (2015) has proposed a multidimensional framework for evaluating the degree to which an external tool is genuinely integrated into a cognitive system, based on criteria including information flow (is it bidirectional?), reliability, durability, trust, procedural transparency, and informational transformation. AI-assisted cognition scores high on some of these dimensions (information flow, reliability, trust) but low on others, particularly procedural transparency. The user typically has no insight into how the AI arrived at its output, which makes it difficult to evaluate whether the output is an extension of one's own reasoning or a replacement for it. This asymmetry matters. A cognitive extension that the user cannot inspect or evaluate is qualitatively different from one that the user understands and controls.

Third, and most fundamentally, cognitive sovereignty is not about the location of cognitive processes but about the authorship of cognitive outcomes and the maintenance of the capacity to produce those outcomes independently. Even if we grant that thinking with AI constitutes extended cognition in Clark and Chalmers's sense, the question remains: can the person still think without the extension? A person who uses a notebook to support their memory retains the underlying capacity for memory formation; the notebook augments without displacing. But a person who routinely uses AI to construct arguments may find that the underlying capacity for independent argument construction has atrophied, a possibility supported by a growing body of empirical evidence reviewed in Section 4. The extended mind thesis tells us that using tools to think is legitimate. It does not tell us that it is safe to lose the ability to think without them. Cognitive sovereignty tracks this distinction. It concerns whether a person retains the capacity for independent cognitive performance and the awareness to monitor whether that capacity is being maintained or eroded, regardless of what tools they use.

Pritchard (2010) has argued that the extended cognition framework creates difficulties for epistemology precisely because it blurs the line between the agent's epistemic contribution and the tool's. Cognitive sovereignty can be understood as the capacity that preserves this line by maintaining the agent's awareness of where their own

epistemic contribution begins and ends, even (and especially) when external tools are deeply integrated into their cognitive practice.

3 3. Distinguishing Cognitive Sovereignty from Adjacent Constructs

Several existing psychological constructs share conceptual territory with cognitive sovereignty. I examine each in turn, identifying both the overlap and the gap that cognitive sovereignty is designed to fill.

3.1 3.1 Critical Thinking

Critical thinking, as traditionally defined (Ennis, 1985; Facione, 1990; Halpern, 1998), refers to the ability to analyze arguments, evaluate evidence, detect fallacies, and reason logically. It is a set of skills and dispositions that can be assessed independently of the tools or technologies involved in reasoning.

Cognitive sovereignty overlaps with critical thinking in requiring evaluative judgment, but it differs in two important respects. First, critical thinking is typically assessed on the quality of reasoning outcomes: can the person identify a logical fallacy? Can they evaluate the strength of evidence? Cognitive sovereignty is concerned with something prior: is the person doing the reasoning at all, or has the reasoning been done for them? A person might retain the ability to identify a fallacy when presented with one but may have lost the habit of constructing arguments independently.

Second, critical thinking does not typically address the problem of source confusion, the tendency to lose track of whether a conclusion was reached through one's own reasoning or absorbed from an external source. This is a central concern of cognitive sovereignty. When AI produces an argument that aligns with one's views and is expressed in a voice that feels like one's own, the critical thinking question (Is this argument valid?) is not the same as the cognitive sovereignty question (Did I produce this argument, or am I just approving it?).

3.2 3.2 Metacognition

Metacognition refers to knowledge about and regulation of one's own cognitive processes (Flavell, 1979; Nelson & Narens, 1990). It includes metacognitive knowledge (understanding one's own strengths and limitations), metacognitive monitoring (tracking how well one is performing a cognitive task), and metacognitive control (adjusting strategies based on monitoring).

Cognitive sovereignty is deeply related to metacognition but is not a subset of it. Metacognition asks: how well am I thinking? Cognitive sovereignty asks: am I the one thinking? Recent empirical work has demonstrated that these are distinct questions. Fernandes et al. (2026) found that participants using AI assistance showed a significant metacognitive calibration gap: they believed they had performed better than they actually had, and, crucially, those with greater technical understanding of AI systems showed larger calibration gaps, not smaller ones. This finding echoes the Dunning-Kruger effect (Kruger & Dunning, 1999), in which people who perform poorly on a task tend to overestimate their performance precisely because they lack the competence to recognize their own deficits, but with a critical twist: in the original formulation, metacognitive deficits were associated with low competence. In the AI-assisted condition, metacognitive miscalibration was associated with higher AI literacy, suggesting that knowing how AI works does not protect against the illusion of having done the thinking oneself, and may, paradoxically, amplify it. A person can have excellent metacognitive monitoring in general while having poor cognitive sovereignty.

3.3 3.3 Intellectual Autonomy

Intellectual autonomy, as developed in epistemology and education (Siegel, 1988; Baehr, 2011; Roberts & Wood, 2007), refers to the disposition to think for oneself, to resist conformity pressures, and to form beliefs on the basis of one's own assessment of reasons and evidence.

Cognitive sovereignty shares intellectual autonomy's concern with self-directed thinking but adds a dimension that the traditional construct does not address: the problem of invisible displacement. Intellectual autonomy is typically threatened by visible pressures, such as peer conformity, authority figures, and propaganda. The person knows they are facing external pressure and can, in principle, resist it. AI-assisted cognition is different. Because the AI's output arrives in response to the user's own prompt, is tailored to their context, and is expressed in natural language that may echo their own voice, the displacement is not experienced as external pressure. It is experienced as assistance, even as one's own thought. Cognitive sovereignty names the capacity to detect this invisible displacement, which intellectual autonomy, as traditionally conceived, does not require.

3.4 3.4 Epistemic Agency

Epistemic agency (Elgin, 2013; Hookway, 2010) refers to the capacity to act as an agent in the formation of one's own beliefs, to seek evidence, evaluate sources, and take responsibility for what one believes. It is closely related to cognitive sovereignty but

differs in emphasis.

Epistemic agency focuses on the actions one takes in the pursuit of knowledge: investigating, questioning, and verifying. Cognitive sovereignty focuses on whether one maintains an active, first-person engagement with the reasoning process through which understanding is formed, and on the capacity to distinguish between genuine and simulated authorship of cognitive outcomes. A person might demonstrate epistemic agency by prompting an AI system with sophisticated questions ('What are the strongest objections to this interpretation of the data?') and critically evaluating its responses; they are, in a sense, actively managing their inquiry. Yet they might simultaneously lack cognitive sovereignty if they could not have produced the reasoning without the system's assistance, if they cannot identify which conclusions they arrived at independently and which they adopted from the AI's output, or if their confidence in their understanding reflects the coherence and fluency of what they received rather than the depth of their own engagement with the problem.

3.5 3.5 Cognitive Offloading

Cognitive offloading (Risko & Gilbert, 2016) refers to the use of external tools or strategies to reduce cognitive demand. It is a well-studied phenomenon with a long history. Writing is a form of cognitive offloading, as is using a calculator or looking up a phone number rather than memorizing it.

Cognitive sovereignty is not opposed to cognitive offloading per se. Some forms of offloading are entirely compatible with cognitive sovereignty: using a calculator to perform arithmetic does not typically erode one's mathematical reasoning, and making notes on paper so as not to forget does not typically undermine one's capacity to think. What makes AI-assisted cognition different is that the offloaded tasks are often constitutive of the cognitive capacity in question. When a person uses a calculator during a complex analysis, the interpretive work, deciding what to calculate, why it matters, and what the result means, remains theirs. The calculator handles a step within a larger cognitive process the person is directing. When a person offloads the task of constructing an argument to an AI, the relationship reverses: the reasoning, the weighing of evidence, and the forming of conclusions, which is the cognitive process itself, is what gets handed over. The same applies when a person asks an AI to summarize a complex topic, synthesize a body of research, or explain a difficult concept: the understanding that would normally emerge from the effort of reading, comparing, and making sense of the material is replaced by a ready-made product that arrives without the process. Over time, the capacity that was offloaded, for instance, the ability to construct arguments or synthesize information independently, may weaken, a concern grounded in the well-established principle of use-dependent plasticity and supported

by emerging evidence reviewed in Section 4. Cognitive sovereignty, understood as an ongoing capacity rather than a one-time judgment, is what tracks whether this process has occurred: whether a person who relies on AI for cognitive tasks still retains the ability to perform those tasks without assistance, and whether they are aware of any change in that ability.

Klein and Klein (2025) have recently introduced the term "hollowed mind" to describe a related risk state, in which frictionless AI access enables users to systematically bypass the effortful cognitive processes essential for deep learning, the kind of sustained, generative engagement with material that produces durable understanding and independent judgment. Their framework, which uses the term "cognitive sovereignty" to describe the goal of maintaining intellectual independence, converges with the present account in several respects, particularly in its emphasis on the qualitative difference between offloading procedural tasks and offloading integrative reasoning. The present paper expands on their work by providing a more precise philosophical grounding for the construct, a systematic differentiation from adjacent constructs, and an articulation of the three component capacities (displacement detection, process connection, and authorship discrimination) that give the concept internal structure.

3.6 3.6 Rational Thinking

Stanovich's (2009, 2011) work on rational thinking and dysrationalia provides an additional point of differentiation. Stanovich distinguishes intelligence (the ability to process information efficiently) from rationality (the disposition to form beliefs and make decisions in accordance with epistemic and practical norms). He demonstrates that these are separable: highly intelligent individuals can be systematically irrational.

Cognitive sovereignty introduces a further distinction. A person can be both intelligent and rational, capable of sound reasoning when they engage in it, while lacking the awareness that their reasoning is increasingly being performed by an external system. Consider a policy analyst who uses AI to draft a position paper: she reviews the output, finds the arguments logically sound, the evidence well-organized, and the conclusions defensible, and approves it. She has applied rational judgment to evaluate the product. But she did not produce the reasoning. If asked to construct the same arguments without AI assistance, she may find she cannot, not because she has become less intelligent, but because the capacity to build the argument from scratch has gone unexercised.

It is worth noting that embracing AI-generated reasoning can be entirely rational under certain conditions: when the person has the independent competence to evaluate the output, understands the limitations of the system that produced it, and makes a deliberate decision to adopt it rather than absorbing it by default. The concern cog-

nitive sovereignty raises is not that people accept AI reasoning, but that they do so without awareness of what they are no longer doing for themselves, and without recognizing that the capacity to do it independently may be diminishing.

Stanovich's framework asks whether people reason well when they reason. Cognitive sovereignty asks whether they are reasoning at all, or whether they have unknowingly ceded the reasoning to a system whose outputs they endorse without having produced themselves. Whether rational thinkers are better at maintaining cognitive sovereignty is an empirical question. It is possible that the habits Stanovich identifies as central to rationality, such as the tendency to question one's own assumptions, seek disconfirming evidence, and resist cognitive shortcuts, offer some protection against unreflective reliance on AI. It is also possible that they do not, because AI outputs do not present themselves as cognitive shortcuts. They arrive as fully formed, well-reasoned responses that satisfy precisely the evaluative standards a rational thinker would apply, making it difficult to recognize that the evaluative act has replaced the generative one.

3.7 3.7 Summary of Distinctions

The following table summarizes the distinctive contribution of cognitive sovereignty relative to each adjacent construct.

Construct	**Central Question**	**Gap Cognitive Sovereignty Fills**
Critical Thinking	Is the reasoning sound?	Is the person doing the reasoning?
Metacognition	How well am I thinking?	Am I the author of the thought?
Intellectual Autonomy	Am I resisting external pressure?	Can I detect invisible displacement?
Epistemic Agency	Am I actively pursuing knowledge?	Do I maintain felt connection to the process?
Cognitive Offloading	Am I delegating cognitive tasks?	Has delegation crossed into substitution?
Rational Thinking	Do I reason well when I reason?	Am I reasoning at all, or endorsing?

Table 1: Cognitive sovereignty and adjacent constructs.

It is worth acknowledging that these distinctions are difficult to draw cleanly in actual circumstances, and that the boundaries between cognitive sovereignty and adjacent constructs will often be blurred in real-world AI use. This difficulty, I believe, is not a weakness in the framework. It is part of what the framework is designed to address, because the conditions created by conversational AI make it genuinely hard to tell where one's own cognitive contribution ends and the system's begins, and a con-

struct that did not acknowledge this difficulty would fail to describe the phenomenon accurately.

4 4. Empirical Motivation

This section does not present cognitive sovereignty as empirically validated. No measurement instrument exists for it yet. What it does is review a growing body of evidence that converges on the phenomenon cognitive sovereignty is designed to name, and that existing constructs have not adequately captured.

4.1 4.1 The Metacognitive Calibration Gap

Fernandes et al. (2026) conducted two large-scale studies (N=246, N=452) in which participants used ChatGPT to solve logical reasoning problems from the United States Law School Admission Test. The studies showed that while task performance with AI assistance improved by three points compared to a norm population, participants also overestimated their own performance by an average of four points. The Dunning-Kruger effect, usually observed in these tasks, in which lower-performing individuals overestimate their abilities while higher-performing individuals tend to underestimate theirs, ceased to exist under AI conditions: all participants, regardless of ability, showed significant overconfidence. Most strikingly, higher AI literacy correlated with lower metacognitive accuracy, not higher. Those who knew more about how AI works were more confident but less precise in judging their own performance.

This finding is directly relevant to cognitive sovereignty because it demonstrates that the problem is not a failure of metacognition in the traditional sense (poor self-monitoring) but a failure of authorship discrimination: participants could not distinguish between what they had figured out and what the AI had figured out for them. Their metacognitive machinery was intact; what had broken down was its ability to track the origin of the cognitive output. The fact that AI literacy amplified rather than corrected this effect suggests that knowing how AI works does not translate into the applied capacity to maintain awareness of what it is doing to one's own thinking, which is precisely the gap cognitive sovereignty names.

4.2 4.2 Cognitive Offloading and Reasoning Depth

Several recent studies have documented effects consistent with the displacement cognitive sovereignty is designed to track. Stadler, Bannert, and Sailer (2024) found that students using ChatGPT for scientific inquiry tasks, such as researching a topic, formulating hypotheses, and evaluating evidence, showed reduced depth of reasoning

(students' analyses became more superficial, with less independent evaluation of evidence and greater reliance on the AI's framing of the problem) compared to those working without AI, a pattern they described as "cognitive ease at a cost." Kosmyna et al. (2025) used neuroimaging to show that brain connectivity and executive control measures were weaker during AI-assisted essay writing than during unassisted writing, providing neural evidence that AI assistance reduces the brain's own active involvement in the reasoning process. A study undertaken in several countries (Gerlich, 2025) found a significant negative correlation between frequent AI tool usage and critical thinking abilities, with younger participants showing both higher AI dependence and lower critical thinking scores.

These findings each have methodological limitations that should be acknowledged. The Gerlich study is correlational and cannot establish directionality. The Stadler study examined a specific task in a specific population. The Kosmyna study, while providing valuable neural evidence, involved a relatively constrained writing task. None of them directly measures cognitive sovereignty as defined here. But taken together, they point toward a pattern that existing constructs struggle to describe: AI is not simply helping people think better or worse, it is changing the relationship people have to the thinking itself, reducing the effort, depth, and neural engagement involved in producing cognitive output while leaving the subjective sense of competence intact or inflated. That gap between the actual cognitive process and the subjective sense of having engaged in it is what cognitive sovereignty is designed to capture.

4.3 4.3 The Expertise Asymmetry

A particularly informative line of evidence comes from field studies comparing how novices and experts perform when given access to AI. Dell'Acqua et al. (2026), in a field experiment with 758 management consultants at Boston Consulting Group, found that for tasks within the current capabilities of AI, such as creative ideation, market analysis, and persuasive writing, consultants using GPT-4 produced work that was roughly 30% higher in quality and 25% faster. The gains were largest for consultants who had scored in the bottom half on a baseline assessment. However, for a task that required integrating quantitative data with qualitative interview evidence in ways that fell outside the AI's competence, what the authors call the "jagged frontier" of AI capability, consultants using AI were 19 percentage points less likely to reach the correct conclusion than those working without it. The term "jagged" captures the fact that tasks that appear similar in difficulty to human workers may fall on opposite sides of what AI can do well, making it difficult to know in advance when AI assistance will help and when it will mislead. Similar patterns of uneven AI impact across skill levels have been documented in customer support (Brynjolfsson et al., 2025) and

professional writing tasks (Noy & Zhang, 2023).

This pattern maps onto the cognitive sovereignty framework. For tasks within AI's capabilities, consultants who performed well at baseline maintained process connection: they used AI to generate initial ideas or draft components of their analysis, but continued to shape the strategy, evaluate the reasoning, and refine the output against their own professional judgment. For the task outside the frontier, meaning the business case that required cross-referencing quantitative data with qualitative interview evidence to reach a strategic recommendation, many consultants across all skill levels absorbed the AI's incorrect conclusion without detecting the error, even though the task fell squarely within their professional expertise. They endorsed a well-structured but wrong answer because it was coherent and persuasive. This is a failure of displacement detection and authorship discrimination: the AI was doing the thinking, so to speak, the output looked like sound consulting work, and the consultants could not tell the difference.

4.4 4.4 The Offloading-Atrophy Trajectory

The evidence reviewed so far concerns what happens to reasoning and self-assessment during AI-assisted cognition: whether people maintain awareness of their own cognitive contribution in the moment of use. But cognitive sovereignty also has a temporal dimension. If the capacities it tracks, displacement detection, process connection, and authorship discrimination, depend on the regular exercise of independent reasoning, then habitual delegation to AI may not only compromise cognitive sovereignty in any given interaction but also weaken the underlying capacity over time. This concern is grounded in the well-established principle of use-dependent plasticity whereby cognitive capacities strengthen with use and weaken without it (Draganski et al., 2004; Maguire et al., 2000). When AI routinely handles tasks that previously required independent reasoning, the cognitive infrastructure supporting those capacities receives less exercise.

Direct evidence for this in the AI context is beginning to emerge. Barcaui (2025) found in a randomized controlled trial that undergraduate students with unrestricted ChatGPT access during self-directed study scored significantly lower on a surprise retention test 45 days later (57.5% correct vs. 68.5% for the traditional-methods group, a medium-to-large effect), suggesting that the ease of obtaining answers from AI led students to engage less deeply with the material during study, resulting in weaker memory formation.

Adjacent evidence from the technology domain paints a more complicated picture. Ward et al. (2017) found that the mere presence of a smartphone reduced available working memory capacity, but subsequent replication attempts have produced mixed

results: a direct replication by Ruiz Pardo and Minda (2022) failed to reproduce the effect, Hartmann et al. (2020) found no effect for short-term or prospective memory, and two meta-analyses reached different conclusions, with Böttger, Poschik, and Zierer (2023) finding a significant overall negative effect and Parry (2024) finding that only working memory showed a significant, albeit smaller than originally reported, effect. The often-cited "Google effect" (Sparrow et al., 2011), in which people showed reduced memory encoding for information they expected to be available online, concerns the strategic allocation of memory resources rather than the degradation of reasoning capacity, and should not be overextended to the present argument.

It is important to distinguish between these forms of technological impact and the one cognitive sovereignty is concerned with. A smartphone sitting on a desk may compete for a person's attention, pulling cognitive resources away from the task at hand. But the person is still doing the thinking; the interference is attentional, not substitutive, meaning the smartphone competes for the person's focus but does not perform any cognitive task on their behalf. The reasoning, judgment, and conclusions remain the person's own, even if produced with divided attention or with the aid of apps and tools on the phone that assist with specific subtasks. When AI handles the substantive cognitive work, something categorically different occurs: the person receives a finished product, a summary, an argument, a recommendation, without having performed the reasoning that would normally produce it. The interference is not with attention but with the cognitive process itself. The evidence for this form of impact is in its early days, but it suggests that outsourcing the process has consequences for the durability of what is learned and, potentially, for the capacity to perform the process independently in the future.

4.5 4.5 What the Evidence Suggests

No single study demonstrates the existence of cognitive sovereignty as a measurable construct. What the evidence collectively suggests is that something is happening to human cognition under AI assistance that existing constructs do not adequately describe. Metacognitive monitoring fails in ways that appear to be specific to AI (Fernandes et al., 2026). Depth of reasoning decreases even when the AI-assisted work product appears competent (Stadler et al., 2024). Neural markers of cognitive effort diminish (Kosmyna et al., 2025). Expertise buffers against the effect in ways that appear consistent with the three components of cognitive sovereignty (Dell'Acqua et al., 2026). And the subjective experience of competence does not track the actual cognitive contribution, which is the central problem cognitive sovereignty is designed to name.

The construct's value, at this stage, is not as a validated psychological measure but as a conceptual tool that organizes these disparate findings into a coherent phe-

nomenon and generates specific questions for empirical investigation.

5 5. Empirical Questions the Framework Motivates

If cognitive sovereignty is a real and distinct capacity, several empirical questions follow. I present these not as predictions but as questions the framework makes it possible to ask.

First, is cognitive sovereignty distinct from metacognition? The Fernandes et al. finding that AI literacy amplifies metacognitive miscalibration suggests that the capacity to track the origin and authorship of cognitive output (cognitive sovereignty) may be separable from the capacity to monitor cognitive performance (metacognition). A targeted study would examine whether individuals with strong metacognitive monitoring in non-AI contexts show the same deficits in authorship discrimination when working with AI.

Second, is cognitive sovereignty domain-specific? A person's capacity to maintain cognitive sovereignty in their area of expertise may differ significantly from their capacity in domains where they lack independent knowledge. A physician reviewing an AI-generated differential diagnosis in her specialty, for instance, may maintain strong displacement detection, while the same physician reviewing an AI-generated legal analysis may lack the independent competence to notice when the reasoning has been done for her. The expertise asymmetry documented by Dell'Acqua et al. supports this possibility. If so, cognitive sovereignty may need to be assessed within domains rather than as a general trait, and the appropriate role for AI in a given domain may depend on the user's capacity to maintain sovereignty within it.

Third, does cognitive sovereignty vary as an ongoing state or is it a stable disposition? The construct as defined here, an ongoing capacity that can be maintained or eroded, suggests it is closer to a state than a trait. Longitudinal studies tracking how the three components of cognitive sovereignty change with sustained AI use would clarify this.

Fourth, can cognitive sovereignty be preserved or restored through deliberate practice? If specific practices, such as forming one's own position before consulting AI, or periodically working without AI assistance, protect the three component capacities, this would provide both evidence for the construct's reality and actionable guidance for education and professional development.

Fifth, does the erosion of cognitive sovereignty carry costs for wellbeing? If self-determination theory is correct that autonomy is a basic psychological need, then declines in cognitive sovereignty should be associated with declines in psychological wellbeing, driven by a reduced sense of being the author of one's own understanding.

This prediction tests the motivational significance of the construct.

6 6. Limitations

Several limitations should be acknowledged. First, cognitive sovereignty as defined here is specific to the conditions created by conversational AI, a technology that operates in natural language and produces outputs that activate human conversational instincts. Future technological developments (e.g., brain-computer interfaces, AI systems that operate below the level of conscious awareness) may create new forms of cognitive displacement that require the construct to be extended or revised.

Second, there is a fundamental reflexive challenge: cognitive sovereignty is partly constituted by the ability to notice its own erosion. This means that the most severely affected individuals may be the least able to recognize it, a paradox that parallels Kruger and Dunning's (1999) findings about metacognitive deficits in low-competence individuals. Any future measurement effort will need to address this challenge, likely through behavioral measures that supplement self-report.

Third, the construct as presented may reflect culturally specific assumptions about the value of autonomous thinking. In epistemological traditions that emphasize communal knowledge, deference to authority, or the primacy of collective over individual inquiry, what I describe as cognitive sovereignty may be valued differently or structured differently. Cross-cultural investigation should interrogate the construct's assumptions, not merely export them.

Fourth, I have not addressed the developmental dimension. How cognitive sovereignty develops in children and adolescents who grow up with AI systems as conversational partners, educators, and cognitive assistants is a critical question that requires its own theoretical and empirical treatment. The developmental trajectory of cognitive sovereignty, and whether there are critical periods during which it must be established or risks permanent impairment, is a pressing research priority that the present framework motivates but does not resolve.

Fifth, the relationship between the three components (displacement detection, process connection, and authorship discrimination) is presented here as a conceptual architecture, not an empirical finding. Whether these components are hierarchical, independently variable, or compensatory, whether they share a common underlying factor or represent genuinely distinct capacities, is an open question. The operationalization of cognitive sovereignty will require this question to be addressed, and the answer may well revise the conceptual structure proposed here.

7 7. Conclusion

I have argued that the cognitive consequences of habitual AI use require a construct that existing frameworks do not provide. Critical thinking, metacognition, intellectual autonomy, epistemic agency, cognitive offloading, and rational thinking each capture part of the picture, but none of them names the specific capacity at risk: the capacity to maintain genuine authorship over the process by which one arrives at one's beliefs, judgments, and conclusions, in an environment where the most fluent, responsive, and apparently comprehending conversational partner available is a system that understands nothing, due to its fundamental lack of the experiential and semantic grounding that understanding requires.

Cognitive sovereignty is an attempt to name that capacity precisely enough to investigate it. The three components proposed here, displacement detection, process connection, and authorship discrimination, provide the construct with internal structure. The systematic comparison with adjacent constructs establishes where it overlaps and where it diverges. And the empirical evidence reviewed, while not constituting validation of the construct itself, demonstrates that something is happening to human cognition under AI assistance that existing constructs have not adequately captured.

Whether cognitive sovereignty will survive contact with empirical data, whether its components hold together or require revision, whether it proves to be domain-general or domain-specific, whether it is measurable, these are open questions. But they are questions worth asking. If the construct proves valid and measurable, it would provide the empirical foundation needed to assess the long-term cognitive consequences of AI adoption across educational, professional, and civic domains, consequences that are currently debated on the basis of intuition rather than evidence. If it does not, if it collapses into existing constructs or fails to predict meaningful outcomes, then the concerns motivating this paper, while philosophically coherent, do not warrant the institutional and pedagogical responses they might otherwise demand. Either way, the question is empirical, and answering it requires the kind of conceptual groundwork this paper has attempted to provide.

References

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.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way:

Creating desirable difficulties to enhance learning. In M. A. Gernsbacher et al. (Eds.), *Psychology and the Real World* (pp. 56–64). Worth Publishers.

Brynjolfsson, E., Li, D., & Raymond, L. R. (2025). Generative AI at work. *Quarterly Journal of Economics*, 140(2), 889–942.

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). How people use ChatGPT. NBER Working Paper No. 34255.

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.

De Freitas, J., Oğuz-Uğuralp, Z., & Uğuralp, A. K. (2025). Emotional manipulation by AI companions. Harvard Business School Working Paper No. 26-005.

Dell'Acqua, F., McFowland, E., III, Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, 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*, Articles in Advance, 1–21.

Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.

Draganski, B., Gaser, C., Busch, V., Schuierer, G., Bogdahn, U., & May, A. (2004). Neuroplasticity: Changes in grey matter induced by training. *Nature*, 427(6972), 311–312.

Elgin, C. Z. (2013). Epistemic agency. *Theory and Research in Education*, 11(2), 135–152.

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership*, 43(2), 44–48.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886.

Facione, P. A. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. The Delphi Report. California Academic Press.

Fernandes, D., Villa, S., Nicholls, S., Haavisto, O., Buschek, D., Schmidt, A., Kosch, T., Shen, C., & Welsch, R. (2026). AI makes you smarter but none the wiser: The disconnect between performance and metacognition. *Computers in Human Behavior*, 175, 108779.

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.

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1), 6.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains. *American Psychologist*, 53(4), 449–455.

Hartmann, M., Martarelli, C. S., Reber, T. P., & Rothen, N. (2020). Does a smartphone on the desk drain our brain? No evidence of cognitive costs due to smartphone presence in a short-term and prospective memory task. *Consciousness and Cognition*, 86, 103033.

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.

Hookway, C. (2010). Some varieties of epistemic injustice: Reflections on Fricker. *Episteme*, 7(2), 151–163.

Kant, I. (1784). An answer to the question: What is enlightenment? In *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.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv preprint arXiv:2506.08872.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121–1134.

Lackey, J. (2008). *Learning from Words: Testimony as a Source of Knowledge*. Oxford University Press.

Maguire, E. A., Gadian, D. G., Johnsrude, I. S., Good, C. D., Ashburner, J., Frackowiak, R. S., & Frith, C. D. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences*, 97(8), 4398–4403.

Merleau-Ponty, M. (1945/1962). *Phenomenology of Perception* (C. Smith, Trans.). Routledge.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In *The Psychology of Learning and Motivation* (Vol. 26, pp. 125–173).

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187–192.

Parry, D. A. (2024). Does the mere presence of a smartphone impact cognitive performance? A meta-analysis of the "brain drain effect." *Media Psychology*, 27(5), 737–762.

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.

Ruiz Pardo, A. C., & Minda, J. P. (2022). Reexamining the "brain drain" effect: A replication of Ward et al. (2017). *Acta Psychologica*, 230, 103717.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78.

Ryan, R. M., & Deci, E. L. (2017). *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. Guilford Press.

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.

Stanovich, K. E. (2009). *What Intelligence Tests Miss: The Psychology of Rational Thought*. Yale University Press.

Stanovich, K. E. (2011). *Rationality and the Reflective Mind*. Oxford University Press.

Tomasello, M. (2008). *Origins of Human Communication*. MIT Press.

Ward, A. F., Duke, K., Gneezy, A., & Bos, M. W. (2017). Brain drain: The mere presence of one's own smartphone reduces available cognitive capacity. *Journal of the Association for Consumer Research*, 2(2), 140–154.