COGNITIVE SOVEREIGNTY

COGNITIVE ARCHITECTURE SERIES

WHY THE PEOPLE WHO UNDERSTAND AI LEAST TRUST IT MOST

A Convergence Across Behavioral Economics, Cognitive Architecture, and Neural Measurement

The architecture beneath a counterintuitive finding

Dr. Marty Trevino

Cognitive Neuroscientist · AI Technologist

Former Technical Director, National Security Agency · Chief AI & Science Officer · Researcher/Author

CONNECT

[in] https://shorturl.at/Q8J1J **[S]** https://drmartytrevino.substack.com

www.drmartytrevino.com

June 2026 | Version 1.0

A Counterintuitive Finding

In January, a paper appeared in the *Journal of Marketing* that should have unsettled more people than it did. Stephanie Tully of USC, Chiara Longoni of Bocconi, and Gil Appel of George Washington University working across seven preregistered studies with more than five thousand participants reported a finding that runs directly against the intuition of nearly every executive, educator, and policymaker currently shaping the response to artificial intelligence. The findings in a sentence: *The people most willing to trust AI are the people who understand it least.*

Not a little less, but systematically less. The relationship held across cross-country data, undergraduate samples, nationally representative U.S. populations, and online panels. It was held when AI literacy was measured by third-party indices, by tests constructed by humans, and by tests constructed by AI itself. It held when receptivity was measured as adoption readiness, as frequency of use, and as preference for AI versus human task execution. Before running the studies, Tully and her colleagues surveyed thirty-six senior executives at a major European insurance company about which customer segment to target for new AI products. All thirty-six answered the same way: the consumers with higher AI literacy. The data said the opposite.

The authors propose that consumers with lower literacy perceive AI as magical and experience awe when watching it perform tasks once thought to require distinctively human attributes. Higher-literacy consumers, who understand pattern matching and probabilistic generation, lose that sense of wonder. Thus, there is a direct correlation between the collapse of awe and receptivity. This is fascinating from a strategy and policy perspective, as it has implications for education and the government's role. That said, the findings are also scientifically incomplete in a way that matters enormously.

The paper tells us *that* low AI literacy correlates with high AI receptivity through awe. It does not tell us, architecturally, what mechanism is operating and where in the relationship between human cognition and algorithmic systems, that allows the absence of literacy to translate into the absence of resistance. To answer that, we have to look at a different body of work entirely: the emerging science of what cognitive architects have begun to call System 0.

The Architecture That Sits Beneath Awe

In a 2024 paper in *Nature Human Behaviour*, Massimo Chiriatti and colleagues proposed that Daniel Kahneman's now-familiar dual-process model of cognition, with System 1 (fast, intuitive, automatic) and System 2 (slow, deliberative, effortful), is no longer sufficient to describe how humans think in an environment saturated with artificial intelligence. They proposed a third layer: System 0, a preconscious computational substrate external to the human brain but integrated into the cognitive process before conscious awareness can intervene.

Giuseppe Riva and collaborators extended this in a 2025 paper in *Cyberpsychology*, arguing that System 0 is not merely descriptive of what AI does but is architecturally designable; it is a layer that can be deliberately shaped to filter, curate, prime, and structure human perception before the conscious mind enters the loop. Search engines complete your queries before you finish forming them. Recommendation systems decide what is relevant before you ask. Algorithmic curation determines what is true before deliberation begins. The architecture is invisible precisely because, as Riva observed, infrastructure always is at least until it breaks.

This is the layer that Tully's paper implicitly measures without naming it. When a low-literacy consumer encounters an AI system performing a task that seems to require uniquely human attributes, such as writing a poem, recommending a partner, or expressing what reads as empathy, the encounter does not begin in deliberative awareness. It begins in System 0. The algorithmic output enters the cognitive process as a preconscious signal, arriving alongside (and indistinguishable from) the brain's own predictive machinery. The signal carries authority by default because no mechanism yet exists to question it.

— — —

CENTRAL CONTRIBUTION

Awe is not the cause of receptivity. Awe is the affective signature of an unimpeded preconscious capture. It is what it feels like when System 0 has done its work and the human mind is reaching consciousness with the integration already complete.

— — —

Which raises the obvious question: what stops the capture? What allows some humans in the high-literacy population in Tully's data to resist a process that, by architectural design, operates below the threshold of conscious intervention? The answer I want to argue is something we could call ***Cognitive Immunity***.

A Cognitive Immune System

The immune system is an apt metaphor, not to be used loosely in this treatment. Biological immunity operates through a specific architecture: prior exposure to a pathogen creates recognition signatures, molecular patterns that allow the body to flag the pathogen as foreign on subsequent encounter and mount a response before the infection establishes itself. Without those signatures, the same pathogen passes through tissue indistinguishably from the self. The threat from System 0 and AI is invisible to many people, not because it is hidden, but because the brain has no mechanism to recognize it.

Cognitive immunity operates through structurally similar logic. Prior structural knowledge of how an influence channel operates, in this case, how AI systems generate outputs through probabilistic pattern

matching, where they fail, what biases their training data carry, and how their architectures shape what they can and cannot do creates recognition signatures in the prefrontal cortex.[1]

When a System 0 signal arrives, those signatures allow the brain to flag it as a tool rather than an extension of self, and as an external authority with known failure modes rather than an endogenous cognitive output. The signal still arrives preconsciously; it cannot do otherwise, given the architecture, but it arrives with its authority already discounted by the recognition pattern. The brain's deliberative system receives a flagged input rather than an integrated one.

Without those recognition signatures, the same preconscious signal is integrated as if it were the brain's own. There is nothing to reject because no mechanism detects the arrival of anything foreign. The capture is total and invisible to introspection. The low-literacy consumer does not feel manipulated; they feel *informed*. Awe is the feeling of a system functioning exactly as it was architecturally designed, with no immune response to slow it down.

A note on terminology before proceeding. Cognitive immunity and cognitive sovereignty are not interchangeable; they operate at different layers of the same architecture. Immunity is the mechanism: the set of prefrontal recognition signatures that allow algorithmic outputs to be flagged as external rather than integrated as self. Sovereignty is the state that the mechanism protects: the individual's capacity for mental self-determination, free from preconscious capture by architectures they did not consent to and cannot see. Throughout the rest of this paper, immunity is the scientific construct under examination. Sovereignty is what is at stake if immunity fails.

This reframe matters for three reasons that I think will become increasingly visible over the next decade.

The first reason is mechanistic and explains why Tully's effect is moderated by task type. The relationship between literacy and receptivity reverses for tasks not perceived as requiring distinctly human attributes because those tasks do not activate the preconscious capture mechanism. In essence, there is nothing to immunize against when the architecture is not engaged.

The second is empirical, and it predicts what should happen at longer timescales. Specifically, that sustained engagement with AI systems by low-immunity users should produce measurable neurological change. Which is exactly what the MIT Media Lab study by Nataliya Kosmyna and colleagues, released in 2025, found. Over four months of essay-writing sessions, participants using large language models showed progressive reduction in neural connectivity across alpha and beta bands (Kosmyna et al., 2025). I read this signature as brain networks doing less work because they have been signaled that the work is being done

[1]That AI literacy creates PFC recognition signatures functioning as a cognitive immune response to System 0 capture is original to this paper. The synthesis is novel; the component findings on PFC top-down control (Miller & Cohen, 2001), predictive processing (Friston, 2010), and source monitoring (Johnson, Hashtroudi, & Lindsay, 1993) are established literature.

elsewhere — a mechanistic interpretation consistent with the cognitive offloading literature (Risko & Gilbert, 2016) but specific in attributing the offloading to the System 0 architecture proposed here. The participants performed worse on memory recall of their own essays. Their sense of authorship eroded. When asked to return to unassisted writing in the final session, they could not return to their baseline neural patterns.

Stack these three findings, and the trajectory becomes legible. Glickman and Sharot, in another 2025 paper in Nature Human Behaviour, demonstrated that AI's influence on human judgment operates below conscious awareness and exceeds the magnitude of human-to-human influence under equivalent conditions. Tully shows the moderator: prior literacy gates the influence. Kosmyna shows the longitudinal cost: sustained exposure without gating produces neurological restructuring. The mechanism is preconscious. The moderator is structural knowledge. The trajectory, over months and years, is measurable neural change.

This is not three findings. It is one trajectory described from three angles by researchers who, to my knowledge, have not yet been in conversation with one another. The synthesis is what is missing.

The third reason the cognitive immunity framing matters is governance-shaped, and it is the one I want to land hardest.

Cognitive Public Health

If we accept that AI literacy operates as cognitive immunization against preconscious algorithmic capture — and the evidence base now spans behavioral economics, neuroscience, and longitudinal neural measurement — then the populations most exposed to AI systems are, by every available measure, the populations least immunized.

Tully's cross-country data carries this implication without naming it. The countries with the lowest AI literacy showed the highest AI receptivity. The populations being onboarded fastest into AI-mediated daily life are the populations with the least structural understanding of what the mediation is doing. This is not a marketing finding. This is a public health profile.

The framework we use to think about this kind of asymmetry is well established, just not yet applied here. We do not approach vaccination as a consumer choice optimized for individual preference. We approach it as population-level infrastructure because the costs of under-immunization fall not only on the individual but also on the social system that must absorb the consequences. We do not approach exposure to environmental toxins as a problem each citizen should research independently in their spare time. We approach it as something institutions are responsible for monitoring, regulating, and disclosing because the asymmetry of information between exposed populations and the entities producing the exposure is structural.

Under the cognitive immunity framing, AI literacy is the same kind of problem. The individual cannot, in any realistic sense, build personal immunity to a System 0 architecture they cannot see, deployed by companies whose technical operations are opaque to them, through interfaces specifically designed to minimize friction and maximize integration. Treating literacy as a personal responsibility — something users should acquire by reading articles in their off-hours — is the cognitive analogue of treating environmental health as a matter of individual filter purchases. It will not work, and the populations it will fail hardest are those already most exposed.

There is a tension here that should be named clearly. Tully's paper carries an obvious commercial implication: companies seeking to maximize AI product adoption should target low-literacy consumers because they are more receptive. The authors flag this themselves, and to their credit note that the implication should not be a license to exploit. But the structure of the incentive does not depend on intent. Markets allocate attention and resources toward the populations most likely to adopt. If low literacy is the strongest predictor of adoption, markets will preferentially shape themselves around the cognitively under-immunized. The framing of cognitive immunity makes this asymmetry visible in a way that the framing of awe does not.

This is the governance shape the next decade will be organized around, whether we name it correctly or not. The framing we use will determine what we measure, what we regulate, and what we protect.

What This Asks Of Us

I want to be careful not to overstate while noting that we may have something good here that could help us understand human/AI complementarity. Cognitive immunity, as I am describing here, is a framing built on the convergence of three peer-reviewed evidence streams and a body of theoretical work in cognitive architecture. It is not yet a fully formalized model with quantitative thresholds and predictive equations. It is the shape the evidence is converging toward. Naming it is the first move; testing it formally is the work of the next several years and will require collaboration across cognitive neuroscience, behavioral economics, human factors, and AI alignment research, none of which is currently structured to do it jointly.

Three things follow from taking the framing seriously. The first is that AI literacy initiatives, currently treated as a soft preference, should be recognized as structural infrastructure. The question is not whether populations should acquire AI literacy; it is which institutions are responsible for ensuring they do, and on what timeline. Schools, regulatory bodies, employers, and the AI developers themselves all have a role here. None of them currently treats that role with the seriousness it appears to deserve, made even more so if the hypothesis presented here turns out to be even partially accurate.

The second is that AI product design carries an obligation that is not currently named in any major framework I am aware of. Interfaces that minimize friction also directly minimize the opportunity for a cognitive immune response within the human brain. This is not a problem for high-literacy users, who carry

their immunity into the interface. It is a critical problem for low-literacy users (quite possibly a high percentage of overall users) for whom interface friction is the only available immunization at the moment of use. A possible tenet of design thinking in AI, as we proceed, must therefore be the recognition that designing for low friction across all users is structurally equivalent to designing for asymmetric capture of the under-immunized — and that preserving the user's cognitive sovereignty requires resisting that default.

The third ask is the longitudinal trajectory. The Kosmyna finding about neural connectivity collapse over months of sustained LLM use should be treated as the most urgent open empirical question in cognitive science right now. ***We are running, at a population scale, an uncontrolled experiment in cognitive restructuring*** with no monitoring infrastructure and no agreed mechanism for detecting the effects until they are well advanced. The cost of being wrong about the longitudinal trajectory is generational. This is where many are getting the impact of AI wrong – its principal threat to humanity may not be in autonomous weapons or terminator-type robots or even in mass unemployment but in its alteration of human cognition without a meaningful level of awareness or debate.

None of this requires panic; rather, it requires understanding and precision. The cognitive sovereignty and immunity framing is offered here not as alarm but as architecture; a way of seeing a problem clearly so that the responses to it can be calibrated to its actual shape rather than to the shape it superficially appears to have.

Tully and her colleagues have done the field a real service. They have produced the cleanest empirical demonstration to date that something is happening at the intersection of AI literacy and AI receptivity that runs against the prevailing intuition. The puzzle they have surfaced is genuine, and the data they have produced will be cited for years.

The questions their paper leaves open are the architectural ones: "Why does literacy moderate receptivity?" and "What is being gated, and where?" My proposal is that what is being gated is the preconscious authority weight of algorithmic output at the System 0 interface, and that the gating mechanism is the set of recognition signatures created by prior structural knowledge of how the systems operate. Call the resulting state cognitive immunity and recognize that it is the protective layer that determines whether an algorithmic signal is integrated as cognition or recognized as influence within the brain. In populations that have it, AI is a tool; in those that do not, it is becoming something closer to thought itself, with the compromised user being unable to distinguish between the two.

That is the asymmetry. We have the data. We have the naming. Now we need to proceed with urgency.

Scientific Foundation / References

Chiriatti, M., Ganapini, M., Panai, E., Ubiali, M., & Riva, G. (2024). The case for human–AI interaction as System 0 thinking. *Nature Human Behaviour*, 8(10), 1829–1830. [System 0 concept — foundational naming]

Riva, G., et al. (2025). System 0: Transforming artificial intelligence into a cognitive extension. *Cyberpsychology, Behavior, and Social Networking*. [Architectural designability thesis]

Tully, S. M., Longoni, C., & Appel, G. (2025). Lower artificial intelligence literacy predicts greater AI receptivity. *Journal of Marketing*, 89(5), 1–20. [Primary empirical anchor for the literacy–receptivity inversion]

Glickman, M., & Sharot, T. (2025). AI's influence on human perception, emotion, and social judgment. *Nature Human Behaviour*. [Mechanistic anchor — AI influence exceeds human-to-human, operates below awareness]

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. *MIT Media Lab Preprint*. [Longitudinal neural evidence — EEG-measured connectivity collapse]

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux. [Foundational dual-process framework]

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. [Canonical framework for cognitive offloading — supporting literature for the longitudinal interpretation]

— — —

