Governing Semantic Generation: From Hallucination to Rule Ecologies*

Eric Caplan, PhD
CEO and Founder, GenieCore
A Working Historian's Case for Practical AI Governance

I spent the better part of four years writing a book that included 750 citations. Every quote verified against primary sources. Every page number checked. Every date confirmed. The manuscript, *When Healing Harms*, draws on 120 hours of video testimony from a psychiatric malpractice trial that most scholars had gotten wrong for four decades. I know what it takes to get facts right.[1]

When I finished, I did something unexpected: I designed, developed, and deployed a citation resolution engine. Not because I thought machines could replace the historian's judgment, but because I'd spent enough time watching language models confidently fabricate sources to wonder whether the problem might be solvable. Eight weeks of coding later, I had a working version of GenieCore—a textual resolution engine that does what LLMs cannot (or cannot yet) achieve: identify, verify, and format citations nearly perfectly from "messy citations" and URLs serving as "placeholders" in early drafts.

What I discovered surprised me. The critics who warned about AI "hallucination" were right about the diagnosis. But they'd missed the cure—one that was hiding in plain sight, in the everyday practices of historians, lawyers, doctors, and anyone else who has ever had to get facts straight under conditions of consequence.

The Hallucination Problem, Correctly Stated

* *This essay was conceived and first drafted entirely without AI assistance while walking with Ernie, my Bernedoodle, in the woods. The final version was revised through a rigorous exchange with Claude (Anthropic's Fable model), which proposed edits and suggested quotations for the final version; I accepted, rejected, or rewrote each one. Every quotation and citation in the published version was verified against primary sources before publication — using CitateGenie and QuotateGenie, a citation-resolution engine and a quotation-verification engine I built for precisely this purpose. The ideas, the claims, and the responsibility for every sentence are mine. Generation was assisted; commitment is the author's — which is rather the point.*

[1] Eric Caplan, *When Healing Harms: The Doctor Who Put a Hospital on Trial—and the Case That Shook Psychiatry* (Oakland: University of California Press, 2026).

The dominant framing of AI hallucination treats it as a bug to be squashed. Improve the training data. Refine the loss function. Add more alignment layers. The assumption is that with enough engineering, we can build language models that simply stop making things up.

This assumption is wrong, and the intellectual groundwork for understanding why it's wrong has been available for over a century.

Ferdinand de Saussure, the Swiss linguist whose *Course in General Linguistics* (1916) launched modern semiotics, established that language is a system of arbitrary signs. The word "tree" has no necessary connection to the woody plant it signifies. Meaning emerges from relationships within the system of language, not from reference to the world. As one of Saussure's translators, Roy Harris, put it: "The essential feature of Saussure's linguistic sign is that, being intrinsically arbitrary, it can be identified only by contrast with coexisting signs of the same nature, which together constitute a structured system."[2]

What Saussure showed for human language frames the right worry about language models. An LLM operates entirely within the symbolic system. It is trained on the patterns of language—which words follow which, in what contexts, with what frequencies—not on the relationships that connect those words to the world. Whether today's largest models nonetheless acquire something like those relationships is a genuinely open question, and I will return to it. But Saussure names the essential asymmetry: fluency is a property of the sign system. Truth is not.

Memory as Reconstruction

In 1932, the British psychologist Frederic Bartlett published *Remembering*, one of the most important books ever written about how human cognition actually works. Bartlett asked English university students to read "The War of the Ghosts," a Native American folk tale chosen precisely because its cultural logic would be unfamiliar to his subjects. Then he had them recall the story at intervals—sometimes years later.

What Bartlett found was systematic distortion. His subjects shortened the story. They dropped the supernatural elements that didn't fit their worldview. They changed "canoe" to "boat" and "hunting seals" to "fishing." They didn't do this maliciously or incompetently. They did it because that's how human memory works: it reconstructs, filling gaps with what makes sense given prior expectations. As Bartlett concluded, "Remembering is not the re-excitation of innumerable fixed, lifeless and fragmentary traces. It is an imaginative reconstruction, or

[2]Roy Harris, translator's introduction to Ferdinand de Saussure, *Course in General Linguistics*, trans. Roy Harris (1983; repr., London: Bloomsbury, 2013), x.

construction, built out of the relation of our attitude towards a whole active mass of organised past reactions or experience."[3]

The cognitive science has only deepened since Bartlett. Karl Friston, the neuroscientist whose work on predictive processing has reshaped our understanding of perception, argues that the brain is fundamentally a prediction machine. We don't passively receive sensory input; we actively generate hypotheses about what we're likely to encounter, then update those hypotheses based on prediction error.[4] Anil Seth, author of *Being You: A New Science of Consciousness*, captures this with a striking phrase: perception is "controlled hallucination."[5] We never experience the world as it simply is, Seth argues; what we perceive is the brain's "best guess" of what's out there. Think Plato's cave.

If perception itself is controlled hallucination—top-down prediction constrained by bottom-up sensory feedback—then hallucination isn't a pathology. It's the baseline condition of cognition. The question isn't whether semantic systems hallucinate. They all do, human and artificial alike. The question is what controls the hallucination.

What the Critics Got Right

In 2021, Emily Bender, Timnit Gebru, and their colleagues published "On the Dangers of Stochastic Parrots," a paper that became famous both for its arguments and for the circumstances surrounding its publication—Gebru was fired from Google after refusing to retract it. The paper's central claim deserves to be quoted directly:

> Contrary to how it may seem when we observe its output, an LM is a system for haphazardly stitching together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning: a stochastic parrot.[6]

---

[3]Frederic C. Bartlett, *Remembering: A Study in Experimental and Social Psychology* (1932; repr., Cambridge: Cambridge University Press, 1995), 213.

[4] Karl Friston, "The Free-Energy Principle: A Unified Brain Theory?" *Nature Reviews. Neuroscience* 11, no. 2 (2010): 127-38, https://doi.org/10.1038/nrn2787, accessed June 10, 2026.

[5]Anil K. Seth, *Being You: A New Science of Consciousness* (New York: Dutton, 2021).

[6]Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York: ACM, 2021), 610–23, at 617, https://doi.org/10.1145/3442188.3445922, accessed June 10, 2026.

As an account of how these systems are built and trained, this is accurate. Language models are sophisticated pattern-matchers operating over linguistic form. They have no communicative intent. They build no model of the reader's mental state. They possess no accountability for the accuracy of their outputs. When GPT-5.5, Claude Opus 4.8, and Gemini 3 produce fluent prose about legal precedents, no one stands behind the claims—no license is at stake, no reputation, no oath. The model is predicting what tokens typically follow other tokens in contexts that pattern-match to legal discourse.

Bender and her colleagues also correctly identified the risks: "The tendency of human interlocutors to impute meaning where there is none can mislead both NLP researchers and the general public into taking synthetic text as meaningful."

We are meaning-seeking creatures. When we encounter fluent language, we assume a communicator with intentions behind it. This assumption serves us well—many but not certainly all times—in human-to-human communication. It becomes dangerous when applied to systems that produce fluent language without any communicative intent at all. Or in a world where "alternative facts" have de facto cultural parity with just facts.[7]

The stochastic parrots paper was right about all of this. Its diagnosis remains the clearest articulation of the risk: synthetic text invites a trust it has done nothing to earn.

A Concession—and Why It Doesn't Matter

Honesty requires a complication. Since the stochastic-parrots paper appeared, interpretability research has given reasons to doubt the strongest version of its claim. Probing studies have found that models trained on nothing but the transcripts of board games develop internal representations of the board; the largest contemporary models appear to develop structures that track space, time, and the properties of objects they have never seen. Whether such structures amount to "meaning" or "reference" is a live debate among researchers and philosophers of language, and it is not one a historian is going to settle. Saussure himself would remind us that the question—what connects a sign system to the world?—is harder than either side tends to admit.

Here is the point that matters: nothing in my argument depends on how that debate resolves. Suppose the critics are wrong. Suppose some future model develops internal structures that genuinely track the world. Its outputs would still arrive as unverified commitments—claims released without any process by which a reader could distinguish the tracked from the

---

[7] *"Conway: Press Secretary Gave 'Alternative Facts,'" Meet the Press, NBC News, January 22, 2017, YouTube video,* https://www.youtube.com/watch?v=B3xsgWHdyN4, accessed June 10, 2026.

confabulated. The epistemic status of a claim depends not only on its content, and not only on the cognitive system that produced it, but on the process by which it can be checked. A citation confirmed against CrossRef carries more warrant than an identical string produced by a model—not because the model is a parrot, but because the confirmation has a known error profile and the generation does not.

This is why the governance argument is durable in a way the semantic argument is not. Every capability improvement leaves it untouched. Smarter models will make fewer errors; they will not make their outputs verified. Verification is not a property a claim can possess on its own. It is a relationship between a claim and a process.

What the Critics Missed

But here is where I part company with the critical literature. Having correctly diagnosed the problem, the critics consistently failed to propose workable solutions. Their prescriptions tend toward the cautionary rather than the constructive: better documentation, more careful deployment, awareness of limitations. These are reasonable suggestions, but they leave the core problem untouched.

The core problem is not that language models hallucinate. The core problem is that we deploy language models without subjecting their outputs to the epistemic governance structures that human societies developed precisely to manage "hallucination" in human cognition.

Consider what we already know how to do. Historians verify claims against primary sources. Lawyers cite precedent and expect opposing counsel to challenge misstatements. Doctors check diagnoses against symptoms, test results, and clinical criteria. Scientists demand reproducibility. Journalists require multiple sources. Even everyday fact-checking—the kind you do when you hear a rumor and think "that doesn't sound right"—involves implicit rules: Is this source reliable? Is the claim consistent with what I already know? Does the person making the claim have something to gain by deceiving me?

These are not epistemically mysterious procedures. They are explicit, teachable, and routinely applied by ordinary people in ordinary circumstances. They have been codified in professional standards, legal procedure, scientific method, and journalistic ethics for centuries.

The astonishing thing is not that we lack the means to govern AI-generated claims. The astonishing thing is that we have not applied the means we already possess.

Why Retrieval-Augmented Generation Is Not the Answer

To be fair, the field has not been entirely idle. Since 2021, researchers and engineers have developed Retrieval-Augmented Generation (RAG)—an architecture that retrieves relevant documents from external databases and includes them in the language model's context before generation. The idea is intuitive: if hallucination arises from the model's lack of grounding in authoritative sources, give it authoritative sources to work with.

RAG represents a genuine improvement over pure generative models. It reduces certain categories of hallucination by anchoring generation in retrieved text, and major AI systems now incorporate retrieval capabilities. But it ameliorates symptoms while leaving the underlying architecture untouched.

The issue is structural. In a RAG system, the language model remains the engine of commitment. Retrieval happens upstream; the retrieved documents are fed into the model's context; and then the model generates output in exactly the same way it would without retrieval—by predicting tokens based on statistical patterns. The model decides what to include from the retrieved context, how to phrase it, whether to combine it with other information, and how confidently to assert it. The retrieval step provides better input. It does not provide verification of output.

Consider what happens when a RAG system retrieves a document containing the information needed to answer a query. The model may reproduce that information accurately. Or it may paraphrase it in ways that introduce subtle errors, blend it with training data into false composites, or contradict the retrieved document entirely because the statistical patterns of its training favor the contradiction. Retrieval improves the quality of inputs to generation. It does not govern the outputs of generation.

The contrast with traditional knowledge systems is instructive. When a lawyer cites a case, opposing counsel can check the citation. When a scientist reports a finding, peers can replicate the experiment. When a journalist attributes a statement to a source, editors can verify the attribution. These are output-verification procedures—checks that occur after a claim is made, before it is committed to the record. RAG offers no equivalent: the retrieved documents are hidden from users, the model's use of them is opaque, and the relationship between retrieved content and generated output cannot be audited.

This architecture recapitulates the fundamental error that has plagued AI deployment from the beginning: treating language model outputs as claims to be trusted rather than candidates to be verified.

The Philosopher Who Saw It Coming

Ian Hacking, the Canadian philosopher of science, spent decades studying what he called "making up people"—the way that scientific classifications don't merely describe the world but actively shape it. "People spontaneously come to fit their categories," he observed. Multiple personality, he showed, "as an idea and as a clinical phenomenon was invented around 1875: only one or two possible cases per generation had been recorded before that time, but a whole flock of them came after."[8] The classification created what it purported to describe.

Hacking would later give the feedback mechanism a name—the "looping effect": classification changes the people it classifies, and the changed people force the classification itself to be redrawn.[9]

What makes Hacking's work relevant here is his recognition that knowledge isn't stabilized by better theory alone. It's stabilized by institutional practices—by the rule-governed procedures through which claims are made, challenged, verified, and either accepted or rejected. The authority of psychiatric diagnosis doesn't come from the rightness of the underlying theory. It comes from the institutional apparatus of DSM committees, clinical training, peer review, and legal standards for expert testimony.

Human knowledge production is embedded in what I'll call "rule ecologies"—interlocking systems of constraint that govern when claims may be made, how strongly they may be asserted, what evidence is required, and what consequences follow from error.

We do not extend these rule ecologies to AI systems. We ask language models to produce claims and then release those claims into the world without passing them through the verification procedures we would apply to any human making equivalent claims.

The Technical Difference: Verification, Not Augmentation

Here I must declare the obvious: I built one of these systems, and I have an interest in their adoption. Judge the argument on the practices, which predate my company by several centuries.

The distinction between GenieCore and RAG is not merely conceptual. It manifests in specific architectural choices that determine what the system can and cannot guarantee.

Consider how RAG handles a citation query. It retrieves documents that appear relevant—abstracts from PubMed, snippets from Google Books, passages from a legal database—

---

[8]Ian Hacking, "Making Up People" (1986), repr. in *The Science Studies Reader*, ed. Mario Biagioli (New York: Routledge, 1999), 161–62.

[9]Ian Hacking, "The Looping Effects of Human Kinds," in *Causal Cognition: A Multidisciplinary Debate*, ed. Dan Sperber, David Premack, and Ann James Premack (Oxford: Clarendon Press, 1995).

concatenates them into the model's context, and generates a response. At no point does the pipeline verify that the output accurately reflects the sources. The model might hallucinate a DOI that conforms to the standard pattern but doesn't exist. It might attribute words to an author who never wrote them. It might combine real details from multiple sources into a composite that never existed. Retrieval improved the probability of accuracy. It provided no guarantee.

GenieCore inverts this logic. Instead of augmenting inputs and hoping outputs improve, it treats every citation as a hypothesis to be tested against authoritative sources—and it refuses to commit any claim it cannot verify.

The architecture is deliberately boring, which is the point. The system queries a local cache first, keyed by canonical identifiers (DOI, ISBN, PMID) rather than search strings. This canonical keying means that multiple different queries referencing the same underlying source all resolve to the same cache entry—a cross-query optimization that produces compounding efficiency gains. If the cache misses, the system queries authoritative public databases—CrossRef, OpenAlex, PubMed—in parallel. These are registries, not generative models: their entries were deposited by publishers and curated by institutions, and their error modes are known and correctable. Matching is deterministic, field by field. Titles, authors, years, volumes, pages, and identifiers are parsed into structured objects and normalized to common formats, so that "January 2024," "Jan '24," and "2024-01" are recognized as equivalent. When any tier produces a high-confidence match, higher-latency tiers are never invoked.

An early version of the system included one further tier: when every database failed, it queried multiple language models in parallel and accepted a resolution only if their independently produced answers agreed, field by field, on the critical fields—title, authors, publication year. That consensus mechanism performed far better than unguarded generation, and building it taught me much of what this essay argues. We removed it anyway. However well-guarded, a generative tier reintroduces the very uncertainty the system exists to eliminate: an answer whose provenance is a model's agreement with other models is still an answer without a source. In the current architecture, no generative model sits anywhere in the resolution path. When the authoritative record cannot confirm a citation, the system does not guess. It says so—it returns the claim marked unverified, names the fields that could not be confirmed, and leaves the commitment to a human. The most important output a verification system can produce is "I could not verify this."

This architecture embodies a principle that RAG ignores: the epistemic status of a claim depends not only on its content but on the process by which it was verified. A citation resolved through CrossRef has higher epistemic warrant than one generated by a language model, even if both

produce identical strings. The process matters because it determines what kinds of errors are possible and how they would be caught.

Citations are, I will be the first to concede, the easy case—the one claim-type for which the world already provides registries to check against. That is precisely why I started there. They are the beachhead: the demonstration that governance is possible at all. The harder claim-types—interpretations, attributions, assertions of fact for which no CrossRef exists—are the research agenda this essay proposes, not a refutation of it.

The Generation–Commitment Distinction

One of the most useful distinctions in human cognition is the one between thinking something and saying it. I can entertain hypotheses privately—imagine outlandish scenarios, consider absurd possibilities—without committing to them publicly. The commitment comes later, when I'm ready to assert the thought under conditions of social accountability.

This distinction is fundamental to professional practice. A doctor may consider many possible diagnoses before committing to one in the medical record. A lawyer may imagine many arguments before asserting one in court. A historian may entertain many interpretations before publishing one with footnotes attached.

Language models collapse this distinction entirely. They generate claims and commit to them simultaneously. There is no internal stage at which a candidate output is held for review, checked against verification procedures, and either released or suppressed.

Building that stage back in is not conceptually difficult. It requires only that we treat LLM output as candidate claims rather than final assertions, and route those candidates through the verification procedures appropriate to their claim-types.

Why This Hasn't Been Done

If the solution is so obvious, why hasn't it been implemented?

Part of the answer is commercial. Language models are marketed as assistants—helpful, fluent, immediately responsive. Inserting verification layers introduces latency, reduces apparent confidence, and makes the systems less fun to use. A chatbot that says "I don't know" or "I need to check that" feels less impressive than one that answers immediately, even if the immediate answer is fabricated.

Part of the answer is cultural. The AI research community prizes technical innovation over applied epistemology. A new architecture is publishable; a new verification procedure is not. The

incentives push toward novel capabilities rather than disciplined governance of existing capabilities.

Part of the answer is that RAG offers an appealing middle ground—something that looks like verification without requiring the hard work of actual verification. Retrieval-augmented systems can claim to be "grounded" in authoritative sources without implementing the checks that would make that grounding reliable. The appearance of rigor substitutes for rigor itself.

But part of the answer, I suspect, is that implementing real epistemic constraints would expose an uncomfortable truth: the problem of AI hallucination is not fundamentally different from the problem of human bullshit. Humans routinely make claims they haven't verified, with confidence that exceeds their evidence, for social and professional advantage. The rule ecologies that govern human knowledge production exist precisely because the default human tendency is to confabulate.

To demand epistemic discipline from AI systems is implicitly to demand it from ourselves. This may be the deepest source of resistance.

The Rule Ecology Framework

Let me be more specific about what a rule ecology for AI would look like. The components are familiar from human practice:

Domain Declaration. What kind of claim is this? Factual, interpretive, speculative, hypothetical? Different claim-types carry different epistemic burdens.

Source Hierarchy. What counts as evidence for this claim? Primary sources trump secondary. Authoritative databases trump web scrapes. Peer-reviewed research trumps blog posts.

Burden of Proof. How strong may the claim be, given the evidence? A verified citation can be asserted confidently. An unverified attribution should be flagged as uncertain.

Consistency Checks. Does the claim contradict other claims already accepted? Contradictions should trigger review, not be papered over with fluent prose.

Temporal Validity. Is the claim still true? Facts change. Sources are updated. What was accurate yesterday may be outdated today.

Genre Recognition. Is this claim being made in a context that permits speculation, or one that demands verification? The rules for brainstorming differ from the rules for publication.

None of these components require new AI capabilities. They require only that we build systems with the discipline to apply the capabilities we already have.

The Philosophical Stakes

The deepest lesson I've drawn from building GenieCore is not technical but philosophical. The question is not "How do we stop AI from hallucinating?" The question is "How do we govern semantic generation under conditions of uncertainty?"

This question applies to human and artificial cognition alike. We hallucinate too. Our memories are reconstructive. Our perceptions are predictive. Our beliefs are shaped by prior expectations as much as by evidence. The difference is that human knowledge production has evolved rule systems—peer review, legal procedure, standards of care, citation practices—that constrain our tendencies toward confabulation.

These rule systems are not perfect. They are biased, contested, unevenly applied, and often captured by existing power structures. But they exist. They function. They provide mechanisms through which claims can be challenged, verified, and either accepted or rejected.

The choice we face with AI is not whether to accept hallucination or eliminate it. The choice is whether to govern it or not.

A Modest Proposal

I am not calling for new regulatory frameworks, though those may eventually be necessary. I am not proposing a technical research agenda, though one is clearly needed. I am making a simpler point.

The tools for governing AI-generated claims already exist. They are the same tools we use to govern human-generated claims. Source verification. Citation checking. Burden of proof allocation. Internal consistency review. These are not exotic technologies. They are practices that every competent researcher, lawyer, doctor, or journalist applies routinely.

What we lack is not capability but will. We have chosen to deploy language models as if they were oracles rather than what they actually are: stochastic pattern-matchers whose outputs require the same verification we would demand of any other source.

Building GenieCore taught me that this verification is not merely possible. It is straightforward. The hard part is not writing the code. The hard part is deciding that verification matters—that we will hold AI systems to the epistemic standards we already know how to apply.

Coda: The Historian's Perspective

I've spent my career studying how institutions manage uncertainty. My first book, *Mind Games*, examined how psychotherapy became established as a legitimate medical practice despite (or because of) the absence of clear theoretical foundations.[10] My forthcoming book, *When Healing Harms*, shows how psychiatric expertise can go wrong, and how legal procedures eventually caught up to expose errors that professional self-governance had concealed for decades.

What these histories share is a recognition that knowledge is not produced by solitary minds pursuing truth. Knowledge is produced by social systems that reward some claims and penalize others, that provide mechanisms for challenge and response, that build verification into the fabric of professional practice.

When we deploy AI systems that bypass these mechanisms—that produce confident claims without verification infrastructure—we are not just risking factual errors. We are undermining the epistemic architecture that makes reliable knowledge possible.

The solution is not to abandon AI. The solution is to embed AI within the rule ecologies we have already built. To require verification before commitment. To treat LLM output as candidates for review rather than assertions to be trusted. To apply to artificial cognition the same disciplines we have learned, over centuries of painful experience, to apply to our own.

Hallucination is not a failure of semantic systems. It is the natural byproduct of meaning-making under uncertainty. Human societies have long managed this risk through rule systems that govern when and how claims may be made. Extending those same systems to AI does not require radical innovation—only conceptual clarity and institutional will.

The problem is not that AI hallucinates. The problem is that we have not yet decided to hold it to the same epistemic standards we already understand and use every day.

Eric Caplan is a historian and the author of When Healing Harms: The Doctor Who Put a Hospital on Trial—and the Case That Shook Psychiatry *(University of California Press, forthcoming 2026). He is the founder of GenieCore, a textual resolution engine and bullshit detector. He can be reached at eric@geniecore.net.*

[10]Eric Caplan, *Mind Games: American Culture and the Birth of Psychotherapy* (Berkeley: University of California Press, 1998).