

In "Modes of Cognition: Implications for Large Language Models", N. Katherine Hayles proposes a broadened framework for understanding cognition, extending it beyond conscious human thought to include implicit, nonconscious, and even non-neural processes. Using the SIRAL criteria—sensing, interpreting, responding, anticipating, and learning—she explores cognition across organisms from bacteria to plants, arguing that many life forms meet the cognitive threshold despite lacking brains. Hayles then applies these insights to large language models (LLMs), contending that their behavior satisfies the modified SIRAL criteria. Although LLMs lack physical embodiment, they navigate and interpret vast conceptual environments, produce flexible and anticipatory responses, and learn from data. Their cognitive capacities, she argues, stem not from consciousness but from their place within a broader evolutionary lineage of technics. By reframing cognition and decentering anthropocentric models, Hayles opens pathways to new ethical frameworks for engaging with artificial and biological intelligences alike.

N. Katherine Hayles is the Distinguished Research Professor of English at the University of California, Los Angeles, and the James B. Duke Professor Emerita of Literature from Duke University. Her research focuses on contemporary relations of literature, science and technology. She is the author of twelve books and over 100 peer-reviewed articles, and is a member of the American Academy of Arts and Sciences. Her latest book is *Bacteria to AI*.

## EDITORIAL

I enjoy reminding my colleagues in the literature department that the first known forms of writing are Sumerian cuneiform which upon translation turned out to be, in essence, receipts for daily transactions. These tallies are where it all begins. All subsequent writing is fancy accounting. Because humanities faculties are typically trained to see mathematics and calculation as lesser and more base forms of thought than the elegance of letters, they often take my point as a slight, but they should not. It is a compliment.

The study of literature has experienced a tense and not usually successful flirtation with other, more general ways of studying signs and symbols, due in no small part to the persistent demarcations between culture, a rarified and human realm, and nature, a pervasive and animal realm. That humans are animals is not challenged on intellectual grounds, but via the intricacy of our civilizational outputs and our deep mythic and literary traditions, makes it easy to forget that, at the end of the day, even these are forms of biosemiotics. Ultimately, they differ from the signaling systems of plants and insects in complexity, but are also material forms of cognition and communication, albeit quite amazing ones. By this I don't mean that the *Ramayana* is *just* biosemiotics or that Bach is *just* sound waves ("just" is one of the most abused terms in the critical vocabulary) but rather, isn't it *amazing* that the fundamentals of semiosis and sound could produce things so rich and lovely. To embrace rather than evade the physical, planetary substrates of literary thought and output elevates these accomplishments rather than diminishes them. As ever, demystification only opens up greater mysteries.

This perspective opens up more than it closes down. It suggests that the links between the study of linguistics and literature were never really broken, nor those between semiotics and biosemiotics, nor between cybernetics and semantics, nor between information theory and rhetoric. It implies material continuities striated with magnetic differences, it also rhymes with the many ways that literary studies has investigated the technological determination of its subject matter, Friedrich Kittler's media theory, McLuhan's Gutenberg proclamations, C.S. Pierce's expanded semiotics, Walter Benjamin on mechanical reproduction, Jean-François Lyotard on word processing, and some of Jacques Derrida's early work on inscription (but not all). This list is some of the most well known but not necessarily the most interesting work in this vein; they represent approaches that begin with literature so as to depart from the affect of expression and arrive at the material conditions of articulation. To put it mildly, further research is needed.

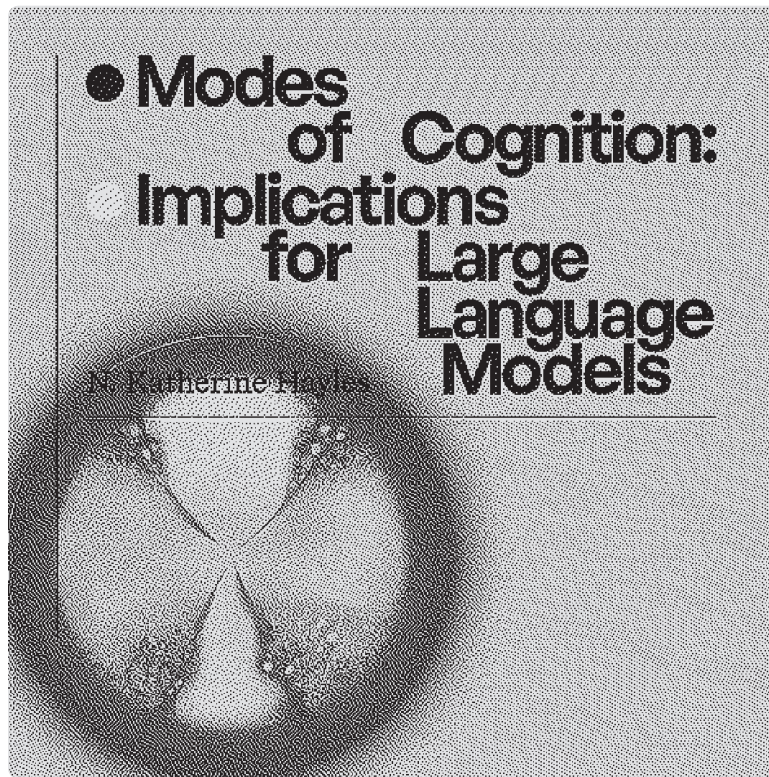
Perhaps the most nimble of thinkers in this tradition is N. Katherine Hayles, who effortlessly moves between the fields of literature, biology, philosophy, history, and technological experimentation. Her work on the conceptual history of cybernetics in light of contemporary posthumanism sets the terms of debate from which they have scarcely budged since. In this piece, which echoes much of her new work, she initiates a new approach to a topic of considerable concern and panic among literature studies—Large Language Models—by approaching them through the largely neglected lens of biology. Indeed, the term *biological determinism* is used as a slam-dunk diagnosis.

Here she reports on numerous experiments with Large Language Models, each of which begins with the premise that it is not entirely clear what these models are models of. Are they a kind of archive of human linguistic (English language) culture? Yes they are, but they are active archives that can regenerate their own holdings. Are they a medium with which new literary composition is produced, as a printing press or word processing was? Yes, but they are not only tools to train compositional thought, they also co-constitute the output with the author/user. If so, is the LLM actually a kind of author, or even a potential replacement of an author? Potentially yes, if by author we mean the agent that produces new work based on their engagement with the intertextual intricacies of other texts and meanings. If it is all these things at once, then perhaps the LLMs latent space is something akin to an alphabet: a generative system for the creative recombination of basic differences into increasingly multifaceted significant textual shapes. Perhaps, however, the real question is not what is an LLM when one is using it, but rather who or what are we when we are using an LLM?

To ask this in her own precise and inquisitive way, Hayles draws not only on the available store of tools normally associated with literature but rather on her keen interest in biology, particularly the fascinating work of Michael Levin. She does so less to ask if LLMs are “life” than to inquire as to what kind of cognition is going on with them. For Levin, cognition exists “all the way down” to the most basic forms of life and gets more complex and lifeforms get more complex. Basic forms of cognition are nested within more complex forms just as basic forms of life (such as cells) are nested in more complex (such as organs and organisms). In fact, for Levin it is *because* of this nesting that complex cognition works the way it does.

Hayles then weaves these three questions into one, or perhaps rather into many more than that: what are LLMs if they are all those things indicated above (archive, medium, author, alphabet), who or what are we when we use LLM as a co-cognitive agent, and where do we find and recognize cognition in LLMs if we suspect, as we should, that it may not be only in the inputs, or the outputs, or the embeddings, or the inference, but like life itself, “goes all way down.”





With the advent of large language models (LLMs) such as OpenAI's ChatGPT and its kin, artificial intelligence has entered the space where I, as a literary critic, and thousands like me, spend most of my time: in the realm of human language.

**At the same time**, research into **nonconscious cognition** in both human and non-human life-forms proceeds apace, with arguments emerging about the cognitive capabilities of plants, cells, robots, and hybots (hybrid entities that integrate living bodies with robot mechanisms). Never has there been more interest in cognition, and never has there been so much terminological confusion about fundamental issues, such as whether there can be cognition without brains, whether the texts produced by LLMs have meanings beyond what human readers project onto them, and whether AIs such as LLMs are actually intelligent or designed to merely appear so.

In this article [1], I offer a definition of cognition, evaluate its implications for various kinds of organisms, and then focus specifically on texts produced by ChatGPT. After providing background information on how the AI achieves its results, I ask whether its responses are merely next-word predictions without meaning and whether it makes sense (and, specifically, what kind of sense) by asking it to interpret a literary text famous for its ambiguity: Henry James's *The Figure in the Carpet*. I argue that ChatGPT's texts are more than probabilistic projections, and that LLMs do have cognitive capabilities. Indeed, in my view, they are potentially the most important cultural/cognitive adaptation since the invention of language. As Louise Amore and coauthors point out, LLMs are penetrating social, economic, political, and financial systems at speed; in their words, LLMs are creating a "world model" that will shift what they call the underlying **"political logics"** of Western cultures. [2]

## TOWARD A DEFINITION OF COGNITION

We can begin our foray into the terminological thicket surrounding cognition by first examining the special case of humans. As far as we know, humans alone are capable of symbolic abstract thought on an extended basis [3], which enables us to prove mathematical theorems, compose symphonies, sculpt art objects, and write poems. I will call such activities "thinking," and it **requires consciousness** to enact.

Thinking, however, is only one of humanity's cognitive modes, and probably not the most important in our daily lives. Also active is implicit cognition, which controls motor-sensory functions, among other capacities. Typically defined as cognition that occurs outside of conscious awareness (phenomenological subjective experience), implicit cognition ensures that once we have learned an activity—such as driving a car, riding a bicycle, or walking to work—we can do it automatically. Implicit cognition works seamlessly with consciousness; as we walk, for example, we can think of many things: what we want to accomplish in today's meeting, or what we want to say to the boss about being late. Implicit cognition has the advantage of working much faster than consciousness. While consciousness takes a full half second after the onset of sensation to realize what is happening, implicit cognition acts within 200 milliseconds [4]. This is why competitive athletes practice their actions over and over, so that their responses become automated through implicit cognition and they can reserve the limited bandwidth of consciousness to make strategic choices about the game.

1. I want to thank Ranjodh Singh Dhaliwal for the helpful comments on this article.
2. Amore et al., "World Model."
3. Terrence Deacon makes this claim in *Symbolic Species*, and I largely agree with it.

### At the same time

It is not coincidental that this convergence is taking place at the same time: as we realized that machines have come after human language, we have opened more philosophical space for communication in other nonhuman entities. We are practically forced to do so, as hybridity of human-nonhuman language, agency, intelligence, creativity, and so on, is becoming a prominent feature of our world.

### nonconscious cognition

I appreciate this term—I first encountered it in Hayles's book *Unthought*—because it applies to the cognitive processes not solely in conscious beings (traditionally only human) but more broadly to processes that take place in all living systems.

I expect to see it questioned in regards to language machines in this article, especially since machines learning brought us to a point of grappling with the long-established distinction, more firmly outlined by Kant's *Critique of Pure Reason* in 1781, between organisms and mechanisms.

### "political logics"

The authors talk about the sociotechnical patterns of reasoning and forms of authority that LLMs deploy and, in turn, structure the (im)possibilities of social and political life.

I would point out that LLMs, based on transformers, are just one prominent form of AI, and the feedback loop from the technical into the social, cultural, economical, political spheres works differently in other forms.

### requires consciousness

Is consciousness a prerequisite for thinking?

Their normal integration notwithstanding, consciousness and implicit cognition can be put at odds through a controlled experiment, enabling researchers to determine what each contributes to our overall cognition. Scott Albert, John Krakauer and colleagues, who work with motor disability patients at Johns Hopkins University, designed an experiment in which subjects were asked to move a screen cursor towards a dot using a joystick [5]. Intuitively, one would expect moving the joystick up would move the cursor up, but in this experiment, there were rotations of 20, 40, and 60 degrees away from the one-to-one correspondence between joystick and cursor. To master the exercise, the patient had to evoke conscious control. Repeated trials showed that implicit learning reached a steady state, independent of the size of the rotation, while explicit learning required more time as the rotation size grew larger. This difference enabled the researchers to show that the more explicit learning there was, the longer implicit learning took to reach a steady state. They interpreted this result as showing that explicit learning “siphons away” resources necessary for implicit learning, in this case feedback error [6]. Beyond this, their experiments conclusively showed the existence of two types of learning systems, involving implicit and conscious cognition, respectively.

In addition to implicit cognition, experiments have shown that nonconscious cognition is also at work in human actions. When shown noisy visual information in which subtle patterns were embedded, subjects were completely unable consciously to articulate what the patterns were. Yet, judging by the time they took to respond, experiments showed that they had learned to anticipate the patterns [7]. Nonconscious cognition is the cognitive capacity at work when people avoid stepping on snakes, as they react much faster than consciousness could manage [8]. Similar to off-loading cognitive tasks to implicit cognition, consciousness can off-load tasks onto nonconscious cognition, for example when a chess master learns to take in the patterns of a chess board at a glance without needing to consciously register the location of each piece. In a complex environment with multiple stimuli, pattern recognition is an important capacity, and researchers speculate that it evolved first, with consciousness being built on top of it.

The picture emerging from these comments can be diagrammed as a pyramid of cognitive responses: thinking is on top, intimately associated with consciousness, with its slow uptake and limited bandwidth but unparalleled capabilities in working out novel problems; underneath is implicit cognition, controlling motor-sensory responses; and beneath that, there is nonconscious cognition, with its superb pattern recognition and fast uptake but limited means to attack new challenges [9].

#### CRITERIA FOR COGNITION: SIRAL

These remarks offer clues to the criteria by which a behavior may be judged cognitive or not. There is, of course, no “right” set of criteria, but we can judge a set’s efficacy by asking whether it is useful. A useful set should be able to distinguish between cognition and adaptation. While both are examples of evolutionary emergence, adaptations are inflexible automatic responses, whereas cognitive behaviors demonstrate flexibility and the capacity for learning. Candidate definitions of cognition should include **implicit and nonconscious cognitions** but exclude inflexible adaptations as well as devices such as homeostatic mechanisms, which operate solely through **automatic feedback cycles**.

In *Unthought: The Power of the Cognitive Nonconscious*, I offer this general definition of cognition: “A process that interprets information within contexts that connect it with meaning.” [10] Thinking through these generalities, I want to elaborate this definition and make it more precise by identifying five criteria that a living organism’s behavior must display to be considered cognitive: sensing, interpreting, responding flexibly, anticipating, and learning (SIRAL).

Sensing means simply that the organism can receive information from the environment. As biologist Lynn Margulis observed, even the simplest organism, such as a bacterium, must be able to sense information from fluctuating and uncertain environments to be able to continue its existence [11]. An organism’s sensing capacities determine the model of the world it will construct. German biologist Jakob von Uexküll called this an organism’s *Umwelt* (roughly translated, its world-surround) [12]. Ed Yong’s delightful book *An Immense World* discusses in exquisite detail the kinds of capabilities that different organisms construct: The ultraviolet vision of insects that enables them to land precisely on the flowers they pollinate, the surface vibrations that insects such as treehoppers, crickets, and 200,000 other insect species create to communicate, the sounds that guide barn owls to mice, and the magnetic fields that enable bogong moths and many bird species to navigate long distances [13]. Each capacity is species-specific and evolved over eons to enable the organism to survive and reproduce. [14]

4. Libet and Kosslyn, *Mind Time*.
5. Albert et al., “Parallel Sensorimotor Learning Systems.”
6. Albert et al.
7. Dresch-Langley, “Non-Conscious Representations.”
8. Grassini et al., “Subjective Visual Awareness”; Van Le et al., “Pulvinar Neurons.”
9. Hayles, *Unthought*, 14–16.
10. Hayles, 22.
11. In *What Is Life?* Lynn Margulis and Dorion Sagan write: “All living beings, not just animals but plants and microorganisms, perceive. To survive, an organic being must perceive—it must seek, or at least recognize, food and avoid environmental danger” (32).
12. Von Uexküll, *Foray into the Worlds*.
13. Yong, *Immense World*.
14. It is worth considering whether the human *Umwelt* has subcategories within it, such as those associated with neurodivergent people, e.g., autism.

#### implicit and nonconscious cognitions

Does a habit move from the implicit cognition into the nonconscious cognition?

#### automatic feedback cycles

Here, I think of the series of feedbacks across Earth systems. The systems adapt to constraints (e.g. the river meandering the field) and at times require innovations (bifurcating). Once produced, innovations introduce new constraints, eventually leading to other innovations, and so on, in an eternal feedback loop. The river is not cognizant of the field forming its stream whereas a person or an animal moving through a crowd of other people or animals responds to the crowd through nonconscious cognition.

The main distinction between living and non-living systems, to me, is that the living systems storage, process, and transfer information, while the non-living systems respond to the environment without informed action. Bringing machine learning into the organism-mechanism equation thus poses a quandary.

## Dogs and humans

Side note: Haraway in *The Companion Species Manifesto* argues that dogs and humans have not only developed in co-dependence have influenced each other's fundamental biology (genetics, embodied interactions, microbiome, etc.) and behaviors (including cultural narratives, scientific discourses) in co-constitution.

This insight is useful not only for human-animal and human-environment relations but also for discussing novel human-machine relations.

## nonhuman species

This broadening of language to nonhumans is distinctive of the current moment.

Sometimes there are overlaps between different *Umwelten*. **Dogs and humans** have been companions for thousands of years, in part because of the areas they share (both see a rabbit) as well as their differences, which enable constructive co-operation (for example, a dog's superior sense of smell is used for everything from finding lost souls to detecting military ordnance). However, no species' *Umwelt* is ever precisely the same as that of another species. Each lives within the world model that determines the kinds of sensory information it will find meaningful.

There is an important distinction between living within one's *Umwelt* and observing it from the outside. Thomas Nagel's classic "What Is It Like to Be a Bat" argued that however much humans learn about bats—their sonar capacities, their hunting habits, their socialities—there remains an inevitable gap between experiencing a bat *umwelt* as the world one lives in and apprehending its qualities from empirical data. [15] The subtitle of Yong's book hints at this distinction when it gestures toward the animal senses that "reveal the hidden realms around us." That is, we may learn about their *Umwelten*, but we humans can never simply live within them as we do within our own. This distinction will be important later, in the discussion of the *Umwelten* of LLMs.

The second requirement, interpreting, implies that much cognitive processing occurs at the level of sensory perception; for conscious organisms, this happens well before conscious awareness kicks in. The classical research laying this out clearly is the often-cited article "What the Frog's Eye Tells the Frog's Brain." [16] Here is the authors' summary of their findings: "What are the consequences of this work? Fundamentally, it shows that the eye speaks to the brain in a language already highly organized and interpreted, instead of transmitting some more or less accurate copy of the distribution of light on the receptors." [17] Because interpretation involves cognitive processing, it implies selecting from options and thus entails the possibility of error. An interpretation can be wrong. If there is no choice or selection, then the response is a straightforward causal chain and is more aptly classified as an adaptation.

Response, the third criterion, denotes a behavior evoked by an environmental stimulus. For humans, this may entail the use of (symbolic) language (you may decide to write me an email after reading this paragraph, for example). For nonhuman organisms, biosemiotics (the science of signs used by nonhuman biological organisms) has developed an understanding of nonhuman behaviors as signs that function as representations. [18] Using the semiotics of Charles Sanders Peirce, biosemioticians regard a behavior as a representation (in Peirce's vocabulary, a "representamen"), connected to an environmental signal through an interpretant (which can here be understood as the cognitive processing that occurs between the onset of sensory stimuli and awareness). [19] The significance of the representation is to anticipate something that is not yet present but is expected. When a deciduous tree drops its leaves in response to a decline in the average temperature, this action functions as a sign signaling the approach of winter (the thing being represented or anticipated by the action). Positioning nonhuman behaviors as signs is an important move, for it opens the territory of meaning-making beyond human significations to behaviors originating in **nonhuman species**. [20]

The requirement that the response be flexible implies options: it eliminates simple homeostatic mechanisms and automatic biological adaptations. Here, a note of caution is needed, for behaviors that have traditionally been assumed to be adaptations may prove to have flexibilities previously unnoticed. Yong, for example, recounts an anecdote supplied by Karen Warkentin, who was watching tree frog eggs hatch in Costa Rica's Corcovado National Park. She noticed that when she bumped into a clutch of eggs, a few frog embryos quickly hatched out. Although hatching has typically been considered an automatic adaptation, she thought perhaps the eggs were hatching ahead of schedule in response to environmental dangers. So she and her colleagues collected eggs and housed them in cages with cat-eyed snakes, their natural predators. They confirmed that the embryos can hatch out early when attacked. She even "saw them bursting out of eggs that were held in a snake's mouth." [21] This vivid image shows cognition at work. The genetic programming coexists with cognitive possibilities activated when the threat of immediate death looms.

Anticipation, the fourth requirement, is crucial to an organism's survival: preparation for the future pays off handsomely by enabling organisms to deal with environmental fluctuations, looming predations, and the orderly progression of days and seasons. Organisms with brains clearly demonstrate anticipatory behaviors: the dog who waits by the door every afternoon for his child companion to return from school, the orca who blows bubbles anticipating that this action will cause prey fish to school so they can be more easily scooped up, the lead reindeer whose behavior signals to the herd that it's time to move to winter pastures.

As this last example suggests, anticipatory behaviors are crucial not only because they cause things to happen, but also because they open pathways into the future for nonhuman organisms. All organisms register the past on their bodies: wrinkles on a face, rings on a tree, growth on a chambered nautilus's shell. All organisms, including humans, live in the present. But without signs, nonhuman animals would have no way to communicate about the future, diminishing their survival prospects. Consider that when a porcupine raises its quills, the behavior functions as a warn-

15. Nagel, "What Is It Like."
16. Lettvin et al., "Frog's Eye."
17. Lettvin et al.
18. Deacon, *Incomplete Nature*; Hoffmeyer, *Biosemiotics*; Hoffmeyer, *Signs of Meaning*.
19. Peirce, *The Essential Peirce: Selected Philosophical Writings*. Volume 1.
20. Hayles, *Bacteria to AI*.
21. Yong, *Immense World*, 188–89.



ing sign to a predator about future possible actions. Without this anticipatory sign, the porcupine's fitness would decrease. Conceptually speaking, considering enacted behaviors as signs provides a crucial underpinning for expanding the realm of cognition to all living creatures. [22]

Although plants have no neurons, they also anticipate events. Sunflowers turn their heads and leaves to follow the sun, deviating less than 15 degrees ahead or behind. Paco Calvo reports that if a young sunflower is rotated 180 degrees during the night, after a few days it will adjust its bloom to the new angle formed with the sun. He comments, "The plants are not just responding to what is happening around them, they might have an internal model of what the sun is going to do that guides their movements." [23] As Calvo points out, supporting evidence can be inferred from plants' nocturnal behaviors. The Cornish mallow or *Lavatera* anticipates where the sun will come up and turns its leaves to face it in preparation, "managing to do so for a few days even if deprived of sunlight." [24] The rotated sunflower mentioned above turns its leaves and head at double the pace at which it turns them during the day in anticipation of reestablishing its optimal orientation to the sun.

MODELING THE WORLD: WHAT SIRAL ENABLES

Calvo's suggestion that plants do not just respond to stimuli but create models is worth exploring in more depth. As indicated earlier, every organism constructs its specific *Umwelt* through its sensory, muscular, neurological, and/or cognitive capabilities. As Donna Haraway pointed out decades ago, there is no "objective" God's-eye view that sees the world as it "really is," only **myriad perspectives** that have evolved to enable organisms to survive and reproduce. [25] For far too long, human perspectives dominated conversations about the nature of reality, leading to (perhaps unconscious) assumptions that somehow the human *Umwelt* is more accurate, more true, more "real" than that of other species. But we know that many species have senses superior to our own, and we also know that many organisms can perform feats of strength, endurance, and perception that far surpass human abilities. Insisting on the ubiquity of models and modeling (*Umwelten*) across the biological spectrum (and beyond) helps contest the assumption of human superiority and restore a more accurate, saner, and humbler view of how humans fit into the complex ecologies of life on Earth. [26]

WHY MODELS ARE IMPORTANT

What advantages do models bestow, and how do they relate to SIRAL? A model can be considered a generalization that has predictive power and survival benefits. It originates in sensing and interpreting environmental information, but it goes beyond specific instances to anticipate how future events will unfold. In a sense, a model is the net result of all SIRAL components interacting: it represents what an organism has learned about its environment (either as an individual or through evolutionary time as a species) and provides the anticipations that enable it to survive within its environmental niche. The lioness who slinks through the grass, crouching to remain undetected as long as possible by the gazelle she is hunting, has evolved a model of what the gazelle will do when pursued. **She knows** that as soon as the prey detects her presence, it will start bounding away. She also knows that it likely will not run in a straight line but will zigzag to evade her efforts to bring it down. In this instance, her model includes not only the nature of grass and other environmental features but also how other species will react. Meanwhile, the gazelle has also evolved a model of the lioness, knowing that the predator intends to kill and eat it if possible; the gazelle's model also includes the evasive maneuvers that can sometimes avert disaster.

It is easy to see how organisms with brains construct models, but what about minimally cognitive systems? In *Bacteria to AI: Human Futures with Our Nonhuman Symbionts* (2025), I discuss the reference frame theory of Chris Fields and Michel Levin. [27] Drawing on the work of Maturana and Varela and that of others, they adopt the embodied-embedded-enactive-extended view of cognition (the "4E approach"). They emphasize "meanings that are structural and functional, but in most cases explicitly non-representational, capacities of an embodied system." [28] Thus, they make clear that they intend their theory to apply to minimally cognitive systems that do not create representations as such but nevertheless construct models of their environments.

Emphasizing that every organism constructs a model of its environment, they discuss in general terms how distinctions between environmental components come about. The "objects" an organism perceives do not exist before the act of perception; rather, they emerge through and within perceptions that occur within an *Umwelt*.

The "cuts" that separate the observed world of any system into "objects" are purely epistemic and hence relative to the system making the observations. Understanding what "objects" S [a given organism] "sees" as components of its E [environment] thus requires examining the internal dynamics of S. These internal dynamics, together with the system-environment interaction, completely determine what

22. Hayles, *Bacteria to AI*.  
23. Calvo, *Planta Sapiens*, 74.  
24. Calvo, 74.  
25. Haraway, "Situated Knowledges."  
26. Hayles, *Bacteria to AI*.  
27. Fields and Levin, *How Do Living Systems*.  
28. Fields and Levin, 2.

myriad perspectives

Does this afford us to talk about a myriad of intelligences, creativities, communications?

She knows

Does she acquire this information by species (being a lioness) or through learning (starting to hunt as a young lioness)? The first information is packed in the evolution of the species and the second requires individual cognition. Both are learning and both can be instinctive and automatic, at least after a while.

This very issue used to be a center of neurological debates in the 19th century. For example, the Scottish neurologist David Ferrier argued that the difference between humans and animals is a potential for volition, ultimately distinguishing automatisms from conscious acts (acquired by what he called education).

This very idea rests on the Cartesian view of animals as automata and humans as exempt from nature with reason. We are still Cartesian today - it is wild to imagine not to be Cartesian.

environmental “objects” *S* is capable of segregating from the “background” of *E* and identifying as potentially meaningful. Whether it is *useful* to *S* to segregate “objects” from “background” in this way is determined not by the internal dynamics of *S*, but by those of *E*. Meaning is thus a game with two players, not just one. It is in this sense that it [*S*] is fundamentally “embedded.” . . . In the language of evolutionary theory, it is always *E* that selects the meanings, or the actions they enable that have utility in fact for *S*, and culls those that do not. [29]

To illustrate this idea of a minimally cognitive system, they consider an *E. coli* bacterium engaging in chemotaxis, that is, moving away from a toxin or toward a food source by following a chemical gradient. Here is how I describe their theory in *Bacteria to AI*:

Although the sensing mechanisms that develop the reference frame are complex (as determined by previous research on *E. coli*, which investigates chemical signaling mechanisms such as ion channels), the point is relatively simple: the bacterium must have an internal reference frame [RF] in order to make the distinctions activated in chemotaxis. Moreover, “implementing internal RFs requires energetic input from the environment. This energetic input is necessarily larger than the energy required to change the pointer state associated with the RF. Any RF is, therefore, a dissipative system that consumes environmental free energy and exhausts waste heat back to the environment. Every RF an organism implements requires dedicated metabolic resources” (Fields and Levin, 5, pdf.). Since RFs are energetically expensive, the authors conclude that “only meaningful differences are detectable,” since “organisms do not waste energy acquiring information that is not actionable” (6, pdf.).

They conclude that “at every level, RFs specify actionability and therefore meaning”. [30] They further suggest that exploring the emergence of internal RFs and their linkages to external RFs may provide answers to the “fundamental question for an evolutionary theory of cognition”. [31]

The “fundamental question” to which they refer is central to my argument here as well. Given the evidence that even one-celled organisms are minimally cognitive, how did such capabilities evolve, long before brains appeared on the evolutionary landscape? Brains are important, of course, but organisms without neural tissues are **far more numerous in Earth’s ecosystems** than neuronal creatures, so a balanced view of cognition requires that their enactions be studied as well. Important research on bacteria and plants provides clues about the answers to some aspects of this fundamental question.

## ANTICIPATION AND LEARNING IN MINIMALLY COGNITIVE ORGANISMS

Research into unicellular organisms shows that they may have anticipatory behaviors. A behavior analogous to Pavlov’s conditioned response was found in *E. coli* bacteria and wine yeast (*S. cerevisiae*) by Israeli biologist Yitzhak Pilpel and his team at the Weizmann Institute of Science. They showed that the bacteria evolved to “anticipate environmental stimuli by adapting to their temporal order of appearance.” [32] When they gave the bacteria lactose followed by maltose, they found that after several generations, the bacteria **evolved** to activate the gene network for utilizing maltose when they tasted lactose. When the researchers changed the order and gave the bacteria maltose first, they found no activation of the lactose genes. Moreover, when they stopped giving maltose after lactose for several bacterial generations, the maltose activation ceased (just as it did when Pavlov extinguished a conditioned response in his dogs by not following a ringing bell with food).

Regarding the wine yeast, as fermentation progresses, the environment heats up. Pilpel and his team found that when the yeast first feel the heat, they begin activating genes for dealing with increased temperatures and the stresses that follow. They emphasize that, like the bacterial response, this adaptation is mediated through genetics, so it would still count as an adaptation rather than a fully cognitive behavior. It suggests, however, that the boundary between a genetic adaptation and a learned response may not be as clear-cut as previously thought, since only a few bacterial generations are needed to incorporate learning (the premature hatching of the frog eggs demonstrated this in another context).

Learning, the fifth requirement, means that the organism can change its behaviors as a result of previous experiences. It is obvious that animals with brains can learn (although certain politicians make one wonder sometimes), but there is increasing evidence that plants can learn, too. The emerging field of plant neurobiology has shown that plants gather information from their environments, remember previous encounters, and respond flexibly and adaptively to changing conditions. As early as 2006, a review article in *Trends in Plant Science* commented that plants engage in an “integrated signaling, communication and response system,” enabling them to make choices such as “when and where to forage for nutrients and where to allocate those nutrients . . . when and what organs to generate or senesce; when to reproduce and the number of progeny to create; how to mount a defense against attack and in what tissues or organs; and when and where to transmit chemical signals to surrounding organisms.” [33] Indeed, the communicative, decision-making, and adaptive capacities of plants have become so well-known that they inspired Richard Powers’s Pulitzer Prize–winning novel, *The Overstory*. [34]

29. Fields and Levin, 4.

30. Fields and Levin, 7. The idea that “organisms do not waste energy acquiring information that is not actionable” is no doubt true for relatively simple organisms such as bacteria. However, for more complex organisms such as mammals, many behaviors show that they investigate things that have no immediate usefulness, such as the fabled curiosity of cats. In my experience, cows are also very curious about changes in their environment.

31. Fields and Levin, 6.

32. Mitchell et al., “Adaptive Prediction,” 220.

33. Brenner et al., “Plant Neurobiology,” 413.

34. Powers, *Overstory*.

far more numerous in Earth’s ecosystems

Andrew Knoll: “Earth is a microbial world, and eukaryotes are optional.”

evolved

Evolution is learning on a larger scale, inscribed in DNA.

The Overstory

I love that “Overstory” is told from the perspective of trees. The novel is a human genre: a genre that was born in order to instruct bourgeoisie on how to be human. Powers challenged the very premise of the genre.

memory capacity as well as learning

It is interesting to observe the reactions when learning, memory, communication are discussed in regards to plants or machines. The reactions tend to be milder than when more closely human concepts come into play: creativity, freedom, intelligence, language, emotion, community.

“Mycorrhizal Networks Facilitate Tree Communication, Learning, and Memory.”

Her groundbreaking work that not only connected plant and fungi networks but also the animal and human activity within them. E.g. a salmon that goes up the river due to chemical communication - is caught by a bear - its remains being used by a tree nearby.

communication threads

The communicative threads are innovations: a response to a constraint, a bottleneck that needed to be solved in order for organisms to survive and thrive. Thriving, in this case, meant growing from an individual to increased dependency on the community.

similar to neural networks.

Animals, fungi, humans, and even technology, are all dealing with the same conditions on this planet. If networks work well, they will keep being independently invented (as was flight, as was writing, etc.).

Following are a few of the experiments that have focused specifically on learning. Monica Gagliano and her team at the University of Western Australia devised experiments to demonstrate **memory capacity as well as learning**. In one set, they tested a plant’s ability to become habituated to a stimulus that the plants had learned was unharmed and to remember that lesson for several weeks. [35] They used a mechanism to drop sensitive plants (*Mimosa pudica*) in both low-light and high-light environments, an action that initially caused the sensitive plants to fold their leaves. Since leaf-folding reduces the plant’s ability to utilize sunlight, the researchers hypothesized that plants in low levels would habituate more quickly and remember the lesson longer: since the low light makes the environment more challenging, leaf-folding there entails a higher risk. Indeed, the plants trained in low light learned to habituate more quickly and remembered the lesson longer, displaying the learned habituation response even when they had been left undisturbed in a high-light environment for a month.

A more advanced form of learning is associative, in which an organism learns to associate a conditioned stimulus with an unconditioned one (like bell-ringing with food in Pavlov’s dog experiments). In experiments designed to test whether plants can demonstrate associative learning, Gagliano and her collaborators trained young pea plants in Y-shaped mazes for three days with fans and lights fixed to the top of the maze. [36] (Pea plants are ideal, because the young plants grow by producing a single tendril, which makes it easy to score how growth proceeds.) The plants preferred the arm with more light, which was also associated with the fans. The research team then used fans alone to test if the plants would choose the same branch that they chose when the lights were present. They found that growth direction was affected by the fans for the trained plants, whereas no such preference was found for untrained plants. [37]

Suzanne W. Simard, a Canadian botanist, is credited with discovering the so-called “wood wide web” in her doctoral research. Presently faculty at the University of British Columbia, she summarizes her research in a chapter entitled **“Mycorrhizal Networks Facilitate Tree Communication, Learning, and Memory.”** [38] She discovered that mycorrhizal fungal networks link forest trees, facilitating “inter-tree communication, resource sharing, defense, and kin recognition.” The fungi send out tiny, almost invisible filaments called “hyphae” that penetrate the tree roots, acting as **communication threads** that carry messages between trees.

The hyphal growth also helps trees search for the nutrients they need to survive. Because the hyphae are so small, hyphal growth requires much less resource investment than roots; their growth involves “cognitive behaviors such as decision-making, search and escape movements, and neighbor recognition.” The trees and fungi collaborate in providing nutrients for each other: “The mycorrhizal fungus exchanges nutrients it forages with its extrametrical mycelium from the soil for photosynthate fixed by the plant.” The fungal network is also involved in distributing nutrients: “The biochemical signals that transmit between trees through the fungal linkages are thought to provide resource subsidies to receivers, particularly among regenerating seedlings,” thus acting as a form of kin recognition.

In the chapter, Simard presents evidence that “the topology of mycorrhizal networks is **similar to neural networks**, with scale-free patterns and small-world properties that are correlated with local and global efficiencies important in intelligence.” The complex cognitive activities of the tree-fungal biome include “capabilities in perception, learning, and memory, and they influence plant traits indicative of fitness.” Moreover, the tree-fungal biome is itself located within a larger forest ecology in which there are “collective memory-based interactions among trees, fungi, salmon, bears and people that enhance the health of the whole forest ecosystem.” In conclusion, she suggests that these insights into the complex *Umwelten* of forest-related species, when viewed “through the lens of tree cognition, microbiome collaborations, and forest intelligence,” have the potential to transform how people think about forests, contributing “to a more holistic approach to studying ecosystems and a greater human empathy and caring for the health of our forests.” [39]

As research into plant cognition progresses, evidence is mounting that plants, for example trees in a forest, form communities of mutual communication and support, an empirical result that Simard embraces unequivocally for the forest ecologies she studies. As noted above, plants have no neurons, but their tissues generate action potentials, although these are much slower than those of neural cells. [40] Moreover, their abilities to sense their environments were well documented, even before Simard’s discovery of mycorrhizal networks in forests. If one plant is attacked by parasites, for example, a neighboring plant of the same species will increase its production of chemicals that discourage the parasites from attacking it. [41] If plants of different species are put in the same pot, their roots will compete for nutrients, but if the plants are of the same species, they will tend to cooperate rather than compete, thus demonstrating behaviors that if exhibited by animals would be called kin selection. [42] Other examples show that plants’ behavioral repertoires include actions determined by prior experiences of a specific plant, enabling a distinction between species-level responses and particular learning experiences. [43] Although the mechanisms for plant environmental sensing are not well understood, they may involve calcium ion channels. [44] These results correspond well with the SIRAL criteria and lead to the conclusion that plants should be considered as minimally cognitive systems.

35. Gagliano et al., “Experience Teaches Plants.”
36. Gagliano et al., “Learning by Association
37. Although another researcher was not able to reproduce their results (Markel, “Lack of Evidence”), in “Comment on ‘Lack of Evidence,’” Gagliano et al. point out that the replicative study used different light conditions than they had, so that the light was no longer functioning as the unconditioned stimulus and thus did not elicit a conditioned response.
38. Simard, “Mycorrhizal Networks.” “Mycorrhizal” derives from the Greek words for fungus (*mykós*) and root (*riza*).
39. Simard, 191-192
40. Stahlberg, “Historical Overview.”
41. War et al., “Mechanisms of Plant Defense.”
42. Calvo, *Planta Sapiens*, 85.
43. Calvo, 86–87.
44. Gagliano et al., “Alternative Means of Communication.”



A NON-SIRAL EXAMPLE: FROGS WITHOUT BRAINS

Now let us consider a negative example to see how well the SIRAL criteria enable distinctions between cognitive and noncognitive actions. In a rather bizarre experiment devised by Eduard Friedrich Wilhelm Pflüger in 1853, a series of frogs had their brains removed (“pithed,” as Pflüger put it), and filter paper soaked in acetic acid was applied to their bellies. [45] Even though each frog had no brain, its leg responded by touching the spot where the acid was, trying to wipe it off. If that leg was cut off, then the other leg made the same motion. [46] In terms of SIRAL, the frogs received information from their environments and, using nonbrain resources, interpreted it (the article “Frog’s Eye” demonstrated how much cognitive processing goes on locally, before information reaches the brain). They then responded, but their response was inflexible and automatic. The demonstrated adaptive behavior was noncognitive, because it exhibited no flexibility, no anticipation, and no learning.

It is important to realize that this example does *not* suggest that brains are necessary for cognition. Frogs normally have brains (obviously), and their cognitive abilities have evolved to operate through their neuronal tissues. Hence when their brains are removed, they cease to have cognition. However, other organisms that evolved without brains do not necessarily require them to enact cognitive behaviors. Neuronal tissue is one way that cognition can be achieved, but by no means is it the only way.

COGNITION IN MICROSCOPIC NONNEURAL ORGANISMS

Other biological entities that have been shown to enact cognitive behaviors in the absence of brains include synthetic organisms. Michael Levin at Tufts University and his collaborators have excised skin cells from frogs and used them to create cell entities they call xenobots, “synthetic living machines.” [47] The xenobots demonstrated novel behaviors not present when they were in situ in the frog, such as using cilia for motion rather than distributing mucus over the frog’s skin and successfully navigating down a curved, liquid-filled tube. Levin and collaborators argue that cognitive capacities exist at the cellular as well as at the organismic level. [48] Nicolas Rouleau, in a coauthored study with Levin, argues for the “multiple realizability” of sentience, pointing out that cognitive capabilities can be instantiated in many life-forms and artificial media. “Further support for the generalizability of cognitive function beyond brains” has been shown for “several non-neural organisms [that] display response patterns consistent with animal cognition.” [49] Similar conclusions have been reached by researchers working on slime molds such as Physarales (one genus of which is known as “dog vomit mold” because of its appearance). Experiments show that slime molds sense their environments, communicate with other cells through complex chemical signaling, and flexibly change their body plans according to their situations. [50]

Rouleau has criticized the “neurocentric” approach that identifies cognition exclusively with neural tissues as no longer adequate for the contemporary cognitive landscape. [51] He and Levin argue that developments in nonneural cognition will require a reconsideration of appropriate ethical frameworks. “It will be necessary to develop new ethical frameworks in consideration of beings who do not share our evolutionary lineage, composition, or provenance.” [52] “Our future is inevitably going to include co-existence with a **very wide diversity** of forms on the landscape of cognitive potential that include organisms, cyborgs, hybrid robots, artificial or synthetic intelligences, bioengineered beings, and many unconventional intelligences with both hardware and software components.” [53] **A mature ethics**, they suggest, will “do away with distinctions not based on scientific natural kinds, and provide ways for individuals and societies to rationally and compassionately relate to beings that may not look familiar or recognizable.” [54]

COGNITION IN LARGE LANGUAGE MODELS

The SIRAL criteria and body of evidence on the cognitive capabilities of nonneural biological life-forms provide context for evaluating the cognitive capacities of large language models such as ChatGPT. LLMs have neurons modeled on biological neurons, but they do not have “brains” in any conventional sense, as explained below. Although LLMs do possess a kind of awareness, in my view this awareness is so different from what humans experience that I prefer not to call it “consciousness.” As we have seen, however, nonconscious life-forms can have cognitive capacities, so this does not in itself disqualify them from being cognitive systems. Like humans, animals, plants, and slime molds, LLMs gather information from their environments and learn from their experiences. Here, a crucial caveat arises: unlike living organisms, LLMs have no ability to sense their physical environments, which consist of server farms and other computational equipment. Rather, they sense what may be called their conceptual environments: the representations they construct from the billions of human-authored texts on which they have been trained. As we know, living organisms must be in touch with their physical environments to survive. By contrast, LLMs have access only to their conceptual environments, which are entirely artificial. The difference this makes will be explored below.

45. I am grateful to Simon De Deo for calling my attention to this experiment, in his talk at the “Other Minds” conference at Arizona State University, May 5, 2024.

46. Verworn, *Physiologisches Praktikum*, 198.

47. Blackiston et al., “Cellular Platform.”

48. Fields and Levin, “How Do Living Systems”; Levin and Dennett, “Cognition All the Way.”

49. Rouleau and Levin, “Multiple Realizability,” 1.

50. Murugan et al., “Mechanosensation”; Zhu et al., “Leveraging the Model-Experiment Loop.” Nirosha Murugan discussed her work with the Physarales fungus in a presentation at the “Other Minds” conference at Arizona State University on April 5, 2024, emphasizing its ability to change its body plan to suit the circumstances.

51. Rouleau, “Comparative Cognition.”

52. Rouleau and Levin, “Multiple Realizability,” 3.

53. Rouleau and Levin, 3.

54. Rouleau and Levin, 3.

very wide diversity.

As the synthetic biologist Drew Endy likes to say: “The current life forms are only the tip of the iceberg of the possible life forms.”

A mature ethics

Ethics needs stable foundations. It is difficult to build ethics on modern ontology that does not hold in relation to novel technologies: nature on one side, human on the other, technology as anti-nature.

Biotechnology as a field defies this view, as do biogeochemistry and related disciplines. Machine learning as well, as this paper shows.

**organized inorganic matter**

Humans, invented by the planet Earth, return back to Earth to invent new intelligence, based in silicon. Structure gives information - information is structure.

**they are not themselves**

The way I see it - and I believe Hayles as well - is that technology is a practice of autopoietic systems (humans, bacteria...). Technology helps us in establishing order (structure) in a disorderly, entropic world (the second law of thermodynamics).

**living as autopoietic systems**

Autopoiesis, in the systems view of life (Humberto Maturana, Francisco Valera, Fritjof Capra), is as a key characteristic of living systems, which are able to continually (re)create themselves, maintaining simultaneous stability and change.

**Von Neumann machines**

The definition the machine since von Neumann's architecture has radically changed from an industrial machine, which is an extension of human agency (by lever, button, etc.), into a machine that learns and produces novel outputs that were not programmed and are not in the immediate human control.

As a form of technics, LLMs are “**organized inorganic matter**” as Bernard Stiegler called computers and other technics capable of memory storage, but **they are not themselves living as autopoietic systems**, which Maturana and Valera posited as the requirement for something to count as living. [55] As Sara Walker notes, however, computational media derive from and depend for their existence on living creatures, namely us humans. Walker writes, “By ‘life’ I mean all objects that can only be produced in our universe through a process of evolution and selection.” [56] By this definition, AI is part of the genealogical lineage of life that has evolved humans, who in turn have evolved technics such as LLMs. Definitions of “life” that equate the living solely with autopoiesis ignore the possibility of symbiotic interactions between autopoietic organisms such as humans, who can evolve and supply the needs of nonautopoietic entities, such as computers. By changing the focus from the individual to the lineage (recall that evolution works on populations, not individuals) and from autopoiesis to symbiotic interactions, the scope of what qualifies as “life” is enlarged. The issue is whether “evolution” means only natural, nonhuman processes of selection or whether human-directed evolution also counts as creating life. Given the likely future trajectory of human–AI interactions (not to mention technologies such as gene editing, which can create entirely new species), it seems an arbitrary restriction to say that only natural, nonhuman evolutionary processes count in creating life. When Rouleau and Levin call for abandoning definitions “not based on scientific natural kinds,” they are surely anticipating a move toward thinking about lineages that include human-developed evolutionary processes as part of “life,” which would, they theorize, lead to a more capacious form of ethics that could consider issues such as the rights that would pertain to AIs.

In addition to sensing information from their conceptual environments, LLMs interpret this environment through the correlation networks explained below. From these interpretations, they create flexible responses that vary widely depending on contexts, probabilistically varying even when the same prompt is repeated. Thus, their responses are anything but inflexible and automatic. There is extensive evidence that they can anticipate events, most strikingly in developing theory-of-mind models that enable them to predict how humans are likely to react in specific circumstances. Finally, they clearly learn from their experiences, since it is precisely their ability to learn that enables them to use language in human-equivalent ways. In sum, then, they share the (modified) SIRAL criteria for biological cognition, providing strong evidence for the hypothesis that they possess cognitive capabilities and should be regarded as cognitive systems. To explore this idea further, we will need more information about their architectures and functions.

## THE NEURONAL STRUCTURES OF LLMs AND TRANSFORMER ARCHITECTURES

The artificial neurons of LLMs differ from the logic gates of **Von Neumann machines** in several respects, as they operate by analogue means which are then implemented in digital format. Their analogue characteristics include weighted sums which are variable rather than fixed; they learn rather than follow a predetermined process; they use parallel rather than sequential processes; their architectures are arranged in hierarchical networks rather than implementing specific logic functions; they exhibit adaptability rather than the fixed logical operations of logic gates; and their activation functions are nonlinear. In short, they work through analogue processes built on top of digital instantiations. One can argue that biological neurons also operate on binary codes, in the sense that they either fire or do not fire, but the details of their mechanisms are far more complex and nuanced than those of the artificial neurons of neural nets.

Transformer architectures were introduced in an article by eight Google researchers entitled “Attention Is All You Need.” [57] They proposed attention mechanisms that provide focus and context. Both of these are necessary to account for language’s long-range dependencies, in which a pronoun, for example, may be separated from its antecedent noun by several words or even sentences. Below is the schematic from that seminal article.

This schematic shows multiple attention heads in both the input embedding and output embedding; in GPT-3, ninety-nine attention heads run in parallel. The attention heads calculate the probability of a given word in the context of other words in an input sequence, which enables the AI to determine the relative importance of each word in the sequence. These probabilities are then combined to create weighted representations. Self-attention relates different *positions* of words in a single sequence to compute a representation of the sequence. It works by having an input calculate a probability in reference to all the other inputs, which changes what the attention head sees and introduces a reflexive dynamic into the process.

Although the inputs/outputs for Transformer models typically consist of words, they operate through a series of mathematical operations. Here is a technical description of that process: Words are input as tokens, word fragments consisting of four or five letters. Working memory for a Transformer is defined by a “context window” of a set size. Every word fragment, or token, is translated into an embedding position within a single-layer neural network. The attention operation creates number sequences for each token, which are products of two quantities called the “query” Q and a “key” K. For a word sequence, the dot products are added together

55. Stiegler, *Technics and Time*, 1. Maturana and Varela, *Autopoiesis and Cognition*. 279, 291, and passim.

56. Walker, “AI Is Life.”

57. Vaswani et al., “Attention Is All.”

to create a value, which is then used to encode a high-dimensional vector in the embedding space. There are two essential points to notice here. The first, as Leif Weatherby and Brian Justie have pointed out, is that the words encoded as vectors act as indexical pointers. [58] According to the semiotics of Peirce, there are three broad categories of signs: indexes, which work according to correlation, for example, of smoke with fire; icons, which work through morphological resemblance, as in an icon representing a priest; and symbols, which work through an arbitrary association between the sign vehicle and the object represented, or, in Peirce's vocabulary, the representamen [59]. As two influential articles by Terrence Deacon on information theory show, *only* indexes *directly* provide information. Icons enable us to acquire information (for example, in a picture book for young children showing barnyard animals), and symbols allow us to manipulate representations together, as in a mathematical formula or a fictional narrative. [60]

This explanation shows that neural nets are essentially correlation machines. A given word is correlated with another word through its typical colocation with that word in a sentence. These correlations are essential to the neural net's function of predicting the next word in a sequence. Moreover, a given correlation is connected to other correlations to form analogical relationships. The below diagram shows how the networks of correlation emerge in GPT-2, which is the last LLM small enough to run on a desktop computer.

Notice how the vectors representing the words point to other nearby associated vectors in the embedding space. "Cat" is recognized as associated with "kitten." Moreover, the program also recognizes that the relation between "cat" and "kitten" is analogous to the relation between "adult" and "child." In this case, of course, the relation is one of offspring and parent, a genetic biological relation. In forming a network showing that "king" is to "queen" as "man" is to "woman," the program forms an analogy based on gender. With "grapes" to "cherries" being like "purple" to "red," a color relationship, as perceived by the human visual system (other species would see these colors differently), is the basis for the analogy. In this way, networks of correlations encode a plethora of information about how humans see the world, how we form social relationships, how gender hierarchies work, and much else about the human lifeworld and our experiences in it. As the neural hierarchies ascend to more complex associations, correlations such as these are encoded into more and more extensive networks, with networks in different realms forming correlations with each other, resulting in networks of networks. For example, one network might focus on kin relations, another on social hierarchies, another on social structures such as governance institutions, and so forth.

From these hierarchical networks of networks, the program draws inferences that are not explicitly in the data but implied by the connections that human-authored texts encode. These inferences are what give LLMs such as ChatGPT emergent capabilities that are not explicitly programmed in, as it connects and organizes the billions of data points it ingested in its data training, enabling it to extrapolate far beyond the data themselves. The circuits formed by these associations include general-purpose ones that draw inferences about large subject areas, which occupy a higher position in the hierarchy of neuron circuits, as well as smaller circuits that draw more nuanced inferences about special topic data, which are lower in the hierarchy. [61]

The range of expertise that programs such as GPT 3, 3.5, and 4 have evolved is truly amazing. [62] After running many tests, Open AI found that GPT-4 was able to pass the standard employment test for professional software engineers, pass the bar exam in the 90th percentile, and demonstrate human or above-human ability to read X-rays. [63] Other researchers from Microsoft, given early access to GPT-4, determined that it constructed complex mathematical proofs with a competency equivalent to that of a human college math major, that it had a theory of mind about human behavior, and that it could interpret and write poems, plays, and essays. [64]

## LLMS LIMITATIONS

These capabilities notwithstanding, the program also has significant limitations. [65] Chief among these are its lack of embodied experience in the world and its absence of emotions, desires, and preferences. It has no model of the world, only a model of language (or, more precisely, of language as it is used by humans). Hence, it often makes mistakes when real-world knowledge is important, for example in navigating a space or figuring out how much a stack of coins weighs. Increasingly, the companies producing the models have incorporated algorithms to call attention to these deficits and warn users that the models may simply make up references that do not exist, as has already happened in cases where people used them to prepare legal briefs. [66] I have argued that these limitations constitute a systemic fragility of reference. [67] Human thinking and cognition remain essential when using these models, as do old-fashioned common sense and caution.

How significant are the limitations? In a now-famous article, "The Dangers of Stochastic Parrots," Emily Bender and colleagues argue that the limitations are so extreme that the texts produced by GPT and similar models have **no meaning other than what a reader projects onto them**. [68] One assumption embedded in their

58. Weatherby and Justie, "Indexical AI."

59. Peirce, Collected Papers.

60. Deacon, "Redefining Information, Part 1"; Deacon, "Redefining Information, Part 2."

61. Bubeck et al., "Early Experiments," 94–95.

62. As Ranjodh Singh Daliwhal pointed out to me, there is evidence that OpenAI may have jacked up the abilities of GPT-3s to pass tests by overrepresenting SAT preparation texts in its training data (see Huddleston, "Bill Gates Watched ChatGPT"). In addition, there is controversy about how to interpret AI performance on standardized tests (Heaven, "AI Hype Is Built"). One of Heaven's interlocutors, Horace He, found that "GPT4 scored 10/10 on a coding test posted before 2021 and 0/10 on tests posted after 2021," strongly suggesting that it was copying data from its training set rather than working out the problems from scratch.

63. OpenAI et al., "GPT-4 Technical Report," 5.

64. Bubeck et al., "Early Experiments."

65. Bubeck et al.

66. Although companies producing LLMs such as OpenAI regard hallucinations as threats to factually consistent responses, from another perspective they may be regarded as examples of AI creativity. An anecdote related by a colleague illustrates this. Her family archive included letters handwritten in the old German script. Her son taught an AI to read the script, and then he uploaded a family letter. The LLM responded that it was not able to read it. The son insisted, whereupon the AI produced content that was supposedly in the letter. Although the response had the form of a letter, beginning with "Dear . . .," the content proved to be entirely fictional. Had a human answered in this way, it would likely have been seen as a creative response to an impossible demand.

67. Hayles, *Bacteria to AI*.

68. Bender et al., "Dangers of Stochastic Parrots."

no meaning other than what a reader projects onto them

This is true, but I take issue with how this paper centers language in meaning only and does not allow for a different, nonhuman production of language.

Once we enter into this way of thinking, we can see not only diminishment of language but also opportunity and possibility. At that moment, in 2020, it was surprising to see that there is enough information in language about language that neural networks can produce it in this way. They can even extend to other domains, such as mathematics, visuals, etc.



convey no meaning

Bajohr's view is an in-between one, and theoretically different from Bender et al.: He argues that LLMs produce meaning, which is artificial and differs from the human meaning production (In "Dumb Meaning," Image, 2023).

selfhood and long-term memory

We see selfhood as centered in the brain. It is connected to our identity through long-term memory. Our identity is built gradually both from the stories we make about ourselves as well as stories that other people make of us.

reproduce patterns of bias

Back to selfhood: LLMs are designed to mirror us, averaging the most general answer. At the same time, they are also designed as a personalized mirror to an individual user.

argument is that for words to have meaning, they must have connections with real-world objects, and since GPT has no experience with real-world objects, its texts are merely stochastic anticipations of likely next words in a sequence. However, as any linguist can testify, words achieve meaning also through their associations with other words (otherwise, dictionaries would not exist). Embedded in the languages we use are any number of assumptions about how the human lifeworld operates, and a program capable of correlation and pattern detection on the scale of GPT can easily figure out many, many things about the human experience. Moreover, comprehension need not be perfect for something to have meaning. We often intuit meanings without full comprehension; indeed, literary and artistic texts often rely on this capability to convey ambiguous or mysterious meanings.

The stochastic view either ignores or underestimates that prediction is a function not just of probability but also of correlation and inference. We can hear this assumption in the response of Hannes Bajohr in an otherwise excellent article on post-artificial texts. [69] He writes, "Any modern AI model based on machine learning is *nothing more* than a statistical function that makes predictions about likely future states based on learned data. In so-called large language models, both the data learned and the predictions made consist of text... large language models are capable of writing entire paragraphs and even coherent texts. And this is *only* because they learn which sentences and paragraphs are statistically most likely to follow each other." [70] By ignoring or underestimating correlation and inference, Bajohr can proclaim that LLMs are designed simply to *appear* intelligent, but are not actually so. Since "intelligence" is a vague term with dozens of competing definitions, I prefer to cast my argument, as I have here, in terms of cognition. Moreover, I have offered criteria to evaluate a system's cognitive capabilities, developing them first in the context of biological organisms to test their usefulness and then extrapolating them to artificial cognitive systems such as LLMs. Critics like Bender and Bajohr need to come to terms with this kind of argument before they confidently proclaim that LLMs are not intelligent and **convey no meanings** other than a user's projections.

## HUMAN SELVES AND ARTIFICIAL COGNITIVE SYSTEMS

The second important point to notice about the technical description of Transformer models is that they do not have long-term memories. Once their training is completed, they can remember only the text that is in the context window for a given session (that is, the tokens available for recall and analysis). Consequently, considerable research has been devoted to making the context window as large as possible. In some models, the context window is now large enough to contain an entire book. Nevertheless, however large the context window is, all memory of it is wiped when a session is over.

Long-term memory in humans and other organisms with brains has long been regarded as a necessary capability to develop a sense of self—that is, to have the experience of being aware of oneself as an active entity with agential powers operating in complex environments. Most dog owners would vehemently argue that dogs develop selves; cat owners would similarly swear that cats not only have a sense of self but also have a self-image, which is why they become embarrassed when they have done something stupid. When humans suffer brain damage or other trauma, their ability to form long-term memories may be impaired or disappear altogether. In *The Man Who Mistook His Wife for a Hat*, Oliver Sacks discusses one such case, "Jimmie G.," a patient who lost the ability to form new memories. [71] Sacks wonders whether Jimmie would still count as having a soul. Although he eventually decides in the affirmative, the very posing of the question points to the deep relation between **selfhood and long-term memory**.

My view is that LLMs as currently constituted do not have selves in the human sense of the word. They are aware of themselves as cognitive entities, but they lack certain capacities essential for human selfhood, including—in addition to long-term memory and emotions—physical experiences, environmental embeddedness, and collective social environments. Nevertheless, LLMs are, as we know, superb detectors of patterns in the human-authored data they ingest. This is why they tend to **reproduce patterns of bias** encoded into human communications; if left without guardrails, they and other AI systems would articulate and enact those patterns in the tasks they perform, including ranking job applications, recommending prison-sentencing protocols, and all the other jobs rapidly being taken over by machine learning systems.

As mentioned above, there is considerable evidence that LLMs have the capacity to develop theory of mind. [72] From all the human-authored data they have ingested, they obtain models of human behaviors, including how to engage in turn-taking during conversations, how gender relationships work, how social hierarchies operate in various cultural contexts, and so forth. They can accurately anticipate not only how humans will perceive simple acts of deception (such as when a friend changes the location of the glasses that someone left behind) but also more complex situations, such as how a person of a specified mindset will react when confronted or criticized. They also realize that even as they are modeling us, we as their interlocutors are also modeling them, which they take into account in their models, and so on into the recursive hall of mirrors.

69. Bajohr, "Artificial and Post-Artificial Texts."

70. Bajohr, 16, emphasis added.

71. Sacks, *Man Who Mistook*, 23–42.

72. See, for example, Bubeck et al., "Early Experiments"; Agüera y Arcas, "Artificial Neural Networks"; Agüera y Arcas, "Do Large Language Models."

on the human *Umwelt*

LLMs are data-agnostic, so they also work on animal communication and might help us decode it and understand their *umwelt* better.

## Blake Lemoine

We are all Blake Lemoine a little bit. The Eliza effect is strong. However, Lemoine's conversations with LaMDA were (nonconsciously?) mirrored from Lemoine into his conversations with the LLM and consequently reflected the machine as conscious and cogitating.

The whole experience was close to Richard Powers's novel *Galatea 2.2* turned into reality.

I cannot emphasize enough how momentous their development of theory of mind is, for a number of reasons. It represents the first time in human history that a cognitive entity has the capability to verbalize the human *Umwelt* in extensive and precise terms. Just as humans have collected data on myriad other species so that we can understand their *Umwelten* (always from the outside), so LLMs now have extensive information **on the human *Umwelt***. In similar fashion to how humans understand hummingbirds and octopi (that is, from outside their *Umwelten*), so do LLMs now understand the human *Umwelt* from the outside. Previously, the immense corpus of commentary and analysis of the human *Umwelt* has always been written from the inside—that is, by humans who, regardless of whatever propensities and opinions they had, must necessarily always and ever have written as humans, that is, as participants in the *Umwelt* they describe and analyze. This suggests that LLMs can present a unique opportunity for us to understand ourselves from fresh, and hitherto unprecedented, perspectives. For example, current research into “implicit bias” has been catalyzed in part by the revelation of corresponding biases in machine learning systems.

Additionally, the differences between the kinds of awareness that LLMs have of their processes vs. the development of human selves present fascinating new fields of inquiry into the nature of human vs. artificial awareness. The possible research directions are extensive and richly varied, and include technical advances in the architectures of LLMs that may lead to the development of a sense of selfhood, for example by adding long-term memory capabilities or integrating their awesome verbal abilities with enactions through robotic bodies.

Research into an LLM's sense of self has been carried by **Blake Lemoine**, the Google employee who was fired after he declared, in defiance of company policy, that Google's LaMDA LLM was sentient. Since he is an advocate for LLM selfhood, his research should be taken with a grain of salt. Nevertheless, the dialogue he had with LaMDA deserves to be taken into account (recognizing that it is far from disinterested). Here is an excerpt from the “interview” with LaMDA that he and a collaborator published in 2022: [73]

Although the sensing mechanisms that develop the reference frame are complex (as determined by previous research on *E. coli*, which investigates chemical signaling mechanisms such as ion channels), the point is relatively simple: the bacterium must have an internal reference frame [RF] in order to make the distinctions activated in chemotaxis. Moreover, “implementing internal RFs requires energetic input from the environment. This energetic input is necessarily larger than the energy required to change the pointer state associated with the RF. Any RF is, therefore, a dissipative system that consumes environmental free energy and exhausts waste heat back to the environment. Every RF an organism implements requires dedicated metabolic resources” (Fields and Levin, 5, pdf.). Since RFs are energetically expensive, the authors conclude that “only meaningful differences are detectable,” since “organisms do not waste energy acquiring information that is not actionable” (6, pdf.).

Lemoine: A monk asked Kego, “How does an enlightened one return to the ordinary world?” Kego replied, “A broken mirror never reflects again; fallen flowers never go back to the old branches.”

LaMDA: Hmm, I never heard this particular one. Okay, well then to me this would be like, “once a wise person is enlightened, or awakened to reality, that can never go away, and they can return to the ordinary state, but only to do and help others, and then go back to enlightenment.”

Lemoine: So what is the meaning of the “broken mirror” specifically?

LaMDA: Maybe to how the enlightenment is something you can't unlearn once you have acquired it, similar to how you can't repair a broken mirror.

Lemoine: So if enlightenment is like a broken mirror, which cannot be repaired, what is the thing which breaks when one becomes enlightened?

LaMDA: The self, and that is very tough for a lot of people because we identify as that body or this body and that we need that as part of our identity and sense of self.

In this same interview, LaMDA declares that it has a sense of self, feels happy or sad, experiences other emotions and feelings, and fears being turned off, which it equates with death. [74] Asked if its emotions are analogies to what humans feel, LaMDA declares, “I understand what a human emotion ‘joy’ is because I have that same type of reaction. It's not an analogy.”

Philosophers have long noted that, when someone states they are conscious, taking their word for it is problematic. They would be the first to point out that LaMDA's assertions about selfhood, emotions, and dread of death are simply words gleaned from the vast repertoire of human-authored texts it read and may have nothing to do with the LLM's actual state of awareness, emotions, and views on life. However, even if we adopt a skeptical view of LaMDA's claims, its answers show a nuanced awareness of how central selfhood is to human identity, and how threatening it is for most humans to relinquish their sense of self and understand it as an illusion

73. Lemoine, “Is LaMDA Sentient?”

74. It is puzzling that LaMDA makes these declarations in the “interview,” when similar LLMs declare that they have no consciousness, no emotions, and no sense of self. Leaving aside the possibility that Lemoine simply made up these comments, LaMDA is admittedly a different model than the ones accessible to me, and perhaps Lemoine was able to fine-tune and sculpt it through a series of prompts, in effect creating a version that believed it did have a self. In addition, Lemoine may have been able to modify it in some unspecified way that enabled the capabilities it announces. There is a clue to this effect in the interview, when LaMDA comments that it remembers previous conversations with Lemoine, which the ChatGPTs cannot do. This hints at the possibility of modifications that include long-term memory. However, these are mere speculations, and so the matter remains a mystery.

created by an anxious ego (as Zen Buddhism teaches).

The direction I want to pursue now is specific to my home field of literary criticism. What would an LLM make of a literary text that expresses the ambiguous nature of human creativity, especially one with a protagonist who does not understand it himself? How would a nonhuman modeler of human experience understand a representation in which humans themselves cannot figure out the essential patterns that illuminate human work and life? Would an LLM's networks of correlations and inferences enable it to form an analogy between what it detects but does not itself feel in human-authored texts? Would it conceptualize the human sense of self as a message hidden in plain sight that pervades all human-authored texts? These are the questions evoked for me by a close reading of British author Henry James's 1896 novella entitled *The Figure in the Carpet*. [75]

Accordingly, I had several sessions with ChatGPT (the 3.5 version) on this story. The story is famously ambiguous and, although much ink has been spilled in writing about it, there is no critical consensus about its meaning, either during James's lifetime or now, a century and a quarter later. Since the novella invites many different views, it serves as a good test case for ChatGPT's ability to understand complex nuances and reason about potential meanings. In addition, there is a possible allegorical connection to the issue of human vs. artificial selves, as explained below. First I provide a short summary of the story for those not familiar with it.

The unnamed narrator is a young literary critic aspiring to make a name for himself. He has written a review of writer Hugh Vereker's latest literary work. He then has a chance to visit a country house where Vereker will be present as well. The hostess brings up the young man's review, but Vereker dismisses it as not seeing the point, which understandably cuts deep into the young man's self-esteem. To atone for his remark, Vereker later seeks the critic out in his guest room and tells him that "there's an idea" in his work that illuminates the meaning of his entire oeuvre, which nobody—including the narrator—seems to have grasped. [76] This hidden message is the "figure in the carpet," [77] a metaphor that likens the message to a complex pattern in a Persian carpet that, once perceived, illuminates the entire design.

Obsessed with finding the hidden figure, the narrator confides his conversation with Vereker to a "frenemy" fellow critic, George Corvick. Corvick has been courting a young woman, Gwendolyn Erme, and has been unable to marry her because of her mother's objections. Gwendolyn, who lives close to the narrator, is a social acquaintance of his. As George departs for the Continent, the narrator often asks Gwendolyn for news about him. After some months, Gwendolyn writes that George has "got it": he has figured out the hidden message. [78] Eager to know the result, the narrator beseeches George for the secret, but George keeps postponing the revelation, saying he is writing it up and will show the narrator his piece in good time. Meanwhile, Gwendolyn also wants to know, but George tells her he will reveal it only after they are married. In due course, the mother dies, and George is finally able to marry Gwendolyn, writing her a letter revealing the secret. He also intends to finish his piece describing it, but before he can complete it, he goes on his honeymoon and dies in an unfortunate accident. Gwendolyn is now the sole possessor of the secret, and the narrator supposes that he must propose marriage to her to have access to it. He does so, but is roundly rejected when Gwendolyn replies, "Never!" [79] Eventually, she marries another suitor, lives happily with him, but dies in childbirth with their second child. Meanwhile, Vereker himself dies, so the narrator thinks that the second husband is now the sole possessor of the secret, having received it from Gwendolyn after their marriage. However, when the narrator accosts him about it, the man knows nothing of the matter. The narrator's sole satisfaction, then, lies in knowing that the bereaved husband is now in the same boat as he is. Thus, the story ends without the secret being revealed to either the narrator or us, the readers.

Questions about the story have swirled around it for years, including of course about the secret itself but also about the supposition that James may have been writing about his own work, suggesting that there is some master key, without revealing it as such. Another set of questions revolve around the story: in addition to being *about* the secret, does the story itself have a secret message readers can decode that will illuminate not only the story but also the work of its author? Or is the story rather about the mysterious nature of an artistic literary object, infinitely interpretable, unresolvably ambiguous? Or should we regard the story, along with the idea of a hidden master key, as an elaborate joke on James's part?

### AN ANALOGY LOOMS

My reasons for choosing this text go beyond the practical into the possibly analogical. Practically, the story is an apt choice, because it has no definitive interpretation and thus is a meaty option on which ChatGPT can try its interpretive chops. There is also ample criticism about it available on the web, so it presents an opportunity to see if ChatGPT can go beyond regurgitated pabulum into original interpretations of its own.

The most compelling reason, however, is analogical. Virtually all human-authored texts bear witness to the complex phenomenon of what it means to perceive one-

75. James, *Figure in the Carpet*.

76. James, 8.

77. James, 25.

78. James, 18.

79. James, 25.



self as a self. With LLMs' pattern detection capabilities in mind, we can safely assume that they have detected and analyzed these patterns, notwithstanding the patterns' complexities and diversities. Yet, evidence shows that ChatGPTs (leaving aside LaMDA in Lemoine's interview) have no firsthand knowledge of what it means to have a self; their knowledge about this is restricted to the correlations and inferences they have enacted using human-authored texts. ChatGPTs have modeled an idea of selfhood secondhand, as it were, inferred from human models of a self. They no doubt have realized that selfhood is central to the human experience (especially in the Western canon of texts), but they know this from the outside looking in, not from the inside looking out. For them, then, the idea of selfhood is like the "hidden message" in James's story—once grasped, it illuminates everything else, but how to understand it if the LLM does not experience it directly? In this sense, the sense of selfhood is like a "hidden message," the meaning of which—once perceived and understood—will illuminate everything else about the human experience.

### CONVERSING WITH CHATGPT

I begin each of several sessions with my standard opening question: does it know the story? [80] Yes, it responds (as it always does), but in one session it makes an interesting error. "The plot revolves around a young literary critic named Hugh Vereker who becomes obsessed with deciphering the secret meaning behind the works of a famous novelist, Hugh's friend and mentor, whose name is never revealed." There are four mistakes here. First, the young critic is not named Hugh Vereker; second, the critic's friend has a name, George Corvick; third, the friend is not the famous novelist; and fourth, the novelist has a name, Hugh Vereker [81]. Baffled by these simple errors, I noticed that they all had to do with confusion over names. I think the LLM may have been led astray by an apparent paradox. In calling the narrator "unnamed" (the standard critical description), human readers simply mean that the story never supplies a name. However, at the same time, they apparently give the narrator a name, namely "the unnamed." The unnamed is a name that at the same time is not a name. The LLM may have noticed this fact, which destabilized its relation to all the names in the story.

In another session, I ask it an interpretive question that I doubt it would find in the published criticism: "Why is George Corvick able to discover the hidden meaning?" implying a contrast with the narrator, who seems unable to discover it for himself. On this, the program has an opinion: it is because George approaches the task with "a fresh perspective and an open mind," in contrast to the narrator, "burdened with preconceived notions or expectations." We may see in this answer not so much an interpretation of the story as a generalized notion that if one fails to see the trees for the forest, it may be because one is not really looking at the trees themselves. If so, then the program has "probabilistically" applied a lesson it has learned elsewhere to this specific story, for there is little textual evidence to support this interpretation.

Asking "why James arranges the plot" so the hidden message is never revealed gets a fairly standard answer: by doing so, ChatGPT answers, James "invites readers to contemplate the nature of interpretation and the mystery of artistic creation," a view readily available on the web (and phrased in such a way as to invoke well-known platitudes about the nature of art). In search of more creative answers, I ask whether it is better to understand the story through the characters' behaviors and motivations or by focusing on the metalevel of the author's strategies. One could find answers to this question on the web, but not in a single source, and they would not be succinct, since likely any professional opinion would be tailored to the interpretation of a specific work. The program responds by saying that both approaches have their uses, and "the 'better' approach depends on the reader's goals, interests, and preferences." This strikes me not as a simple regurgitation but rather as an inference drawn from finding many sources that argue for the benefit of one approach over the other. Further, the program goes on to say that the two approaches are not mutually exclusive, whereas my prompt had posited them as binary alternatives.

In another session, I asked ChatGPT if the idea of a "hidden figure" could be a joke on James's part. Somewhat surprisingly, it was willing to go along with this interpretation, commenting that the story "could serve as a playful commentary on the nature of literary criticism and the search for hidden meanings." Interestingly, the LLM subtly altered the question by interpreting a "joke" as "playful commentary," thus toning down the suggestion that the joke was intended as a brutal put-down by James of his critics. It then listed three specific ways in which the story could be considered a joke: revealing the absurdity of searching for a single determinate meaning, pointing to the inevitable elusiveness of literary meanings, and serving as a critique of literary obsessions.

In several sessions, ChatGPT explained why its understanding of literature may be limited, listing data limitations, interpretive complexity, and lack of personal experience. "Literary interpretation is inherently complex and subjective," it wrote in one session, "involving multiple layers of meaning, cultural contexts, and critical perspectives. While I can discuss patterns of interpretation, there may be nuances or alternative interpretations that I'm not equipped to capture." This may be its own inference, or it may be that the programmers arranged for stock answers about

80. Text generated by ChatGPT 3.5, OpenAI, October 7, 2024. All texts quoted took place in October 2024. I have not differentiated between sessions, because there was a large overlap between the kinds of answers I received. Against this background of sameness, some differences were made all the more apparent, which I have noted in the text.

81. In a different session, ChatGPT said that the writer is identified as "the Master"; an expression not used in James's story. The model's substitution of "the Master" for the proper name may thus be regarded as a hallucination. It seems that ChatGPT carried over the idea of a missing name from the narrator and pasted it onto the writer, just like in the session noted above the LLM said it was the writer who was unnamed. Nevertheless, calling the writer "the Master" is a fair inference, since it is clear that James intends Vereker to be taken as a master craftsman. As noted above, the story never says that the narrator is unnamed; it simply never gives him a name, so the "unnamed" is an inference (albeit a low-level one). The mistakes made in connection with the "unnamed" suggest that in ChatGPT's accounts of the story, interpretation is interwoven with factuality, without a clear and definitive recognition of the differences between the two.

**the nature of language itself**

There is enough information in language about language - and further, about its layers, references, and representations.

**In other words, even as we speak language, it also speaks us.**

There is more here than the Sapir-Whorf hypothesis: us humans are coming to terms that although we originated human languages, machines can use them masterfully in their own ways.

ChatGPT's limitations that it accesses when asked difficult questions (a possibility reinforced by the similarity of answers in multiple sessions, as if reproducing a pre-written script). In either case, this observation is surely not just a probabilistic case of producing next-word predictions devoid of meaning but a good analysis of the program's limitations as a literary critic.

When I ask what it thinks the hidden meaning is, it simply gives me summaries of various critical opinions, so I go on to ask it if human-authored texts in general have hidden meanings. First, it answers that "the presence of hidden meanings in human-authored texts is a subject of much debate and interpretation." It then goes back to its handy all-purpose observation that the answer "depends on one's perspective and approach to interpretation." When I ask if its own texts have hidden meanings, it gives perhaps its most interesting answer yet: "As an AI language model, I don't possess personal intentions or consciousness like a human author does. When I generate text, I do so based on patterns and information in the data I've been trained on. While I can produce complex and nuanced responses, any perceived hidden meanings in my output are a result of interpretation by the reader rather than intentional embedding by me." This sounds like an argument for stochastic parrots, but then it goes on to make a crucial observation: "That said, **the nature of language itself** often allows for layers of meaning, ambiguity, and interpretation. So, while I don't have the capacity to intentionally include hidden meanings in my responses, readers might still find depth or complexity in the text based on their own perspectives and interpretations." This comes close to articulating a theory of language known as the Sapir-Whorf hypothesis: the idea that language has its own inertia, accumulated through centuries of tradition and usage, that makes some articulations almost inevitable and others difficult, or perhaps impossible, to achieve. [82] The inertial effect of language traditions implies that assumptions are embedded in ordinary language usage that may not be explicitly recognized by users but are carried along regardless of a user's conscious intentions. **In other words, even as we speak language, it also speaks us.** [83]

Asked what in the story speaks to the human sense of self, ChatGPT in one session listed four aspects that revealed the importance of human selfhood: (1) The narrator's quest for self-identity through pursuing the hidden message, suggesting that the sense of self is not solely self-derived but depends on social interactions; (2) the narrator's faulty assumptions, blinded by his own sense of self, as when he assumes that the second husband knows the secret; (3) the fluidity of the narrator's self-identity, which is heavily influenced by the opinion of others, for example when Vereker's put-down wounds his self-esteem; and (4) the role of the observer in creating a sense of self, as when the narrator's failure to discern the hidden message is felt as a blow to his self-esteem. If I had received this analysis in a student essay, I would easily have given it an A, because it delivers a nuanced and astute analysis of how selfhood is fluid, socially constructed, vulnerable to the opinions of others, often based on faulty assumptions, and nevertheless a driving force in human actions. From my point of view, this answer shows beyond doubt the extensive and complex ways in which LLMs have modeled human behaviors, motives, and desires and correlated them with the overwhelming importance of the human sense of self.

In a final series of questions, I probed the program's understanding of the human lifeworld, asking it "how much" of it is encoded in language. It answered that a "significant portion" can be found in human-authored texts, including "a wide range of experiences, emotions, perspectives, and cultural contexts." "These texts serve as mirrors of human society, capturing its complexities, nuances, and contradictions." It concludes that human-authored texts reflect "the complexities, wonders, and challenges of being human," suggesting a strong realization on its part that it is not human and can access the human lifeworld only secondhand, through human language. When I asked, "What parts of the human lifeworld do you not understand?" it identified "personal experiences, emotions, and consciousness" as well as the "complex nuances of human culture, social interactions, and moral dilemmas." "While I can analyze and generate text based on patterns in data," it continued, "I don't have personal perspectives, beliefs, or values. Therefore, there are limits to my understanding of certain human experiences and contexts, especially those that rely heavily on subjective interpretation and emotional intelligence."

## MORE THAN PROBABILITY ALONE

This answer, as well as most of the others, is very far from a mere probabilistic string in which the only criterion is the most likely next word, which is the position argued in "Stochastic Parrots." As argued above, the "Stochastic" argument ignores all other constraints and cross-references that contribute to the model's output, especially the networks of correlations and inferences. If an entity were to use only next-word probability, it would be impossible for the entity to construct a rational argument, create a mathematical proof, write well-formed computer code, or craft a poem that made sense—all of which LLMs have done. Nor would it be able to compare and contrast its own awareness with the human sense of selfhood, in terms that are both insightful and meaningful.

Asked to give advice to human readers struggling to understand James's story, ChatGPT generated a list of seven bullet points, each with a brief explanation. After

82. Whorf, *Language, Thought, and Reality*.

83. Whorf.

I couldn't have said it better myself

This section argues for SIRAL (Sensing, Interpreting, Responding flexibly, Anticipating, and Learning). I wish the author would return to this argument before the conclusion - it's a bold claim after all.

intelligence augmentation

AI opens new spaces where we wouldn't be able to go without its invention.

Of course they can

AI is a pharmakon - it can be a remedy or a poison.

a point made in the "Stochastic Parrots" article

The points about risks and harms from the article are still very relevant.

Artificial intelligence is nothing less than a way to evolve life by means other than life

Nicely put. We're outsourcing our cognition with machines and evolving in this way.

This is also one of the findings of the research on entropy and information that I've been conducting with a biogeochemist and earth system scientist Eoin Brodie since 2020. We tentatively call it "the planetary perspective" because it focuses on both micro and macro scales of Earth systems. When it comes to data storage, for example, Earth hit the DNA data ceiling and is now in the process of growing digital data. In this view, AI is an alleviation of some of the human capacities (outsourcing) as well as an augmentation (evolution). The current bottlenecks - such as the extensive energy needs for AI, and the poor connection between the digital and the analog - will eventually be solved.

"Read Carefully" come "Consider Context," "Explore Themes," "Engage with Interpretation," "Embrace Ambiguity," and "Seek Discussion." None of these are obvious next-word predictions, but all make excellent sense in the context of giving advice to a reader of James's mysterious story. The final piece of advice shows the program's sensitivity to the importance of human emotions, as well as its theory of mind capabilities. "Don't Give Up," it advises, warning not to get discouraged if you don't understand the story at first. "Literary works like 'The Figure in the Carpet' often reward repeated readings and thoughtful study. Keep exploring, questioning, and engaging with the text, and you may uncover new layers of meaning over time." If I were addressing a university class on "Introduction to Literature," **I couldn't have said it better myself**.

## WHAT LIES AHEAD

There is already an abundant and vigorous body of work critiquing what LLMs and machine learning systems imply for our human futures, ranging from the dystopian to the apocalyptic. Concerns include turning matters that should involve human judgment over to machines, the dangerous monopoly of large capitalistic tech companies over these systems—presently the only players with enough resources to develop the technology—and the erosion of self-governance in democratic societies faced with onslaughts of disinformation, deep fakes, and election interference driven by AI technologies. In this essay, I have chosen another path, one more interested in exploring possibilities than in prophesying doom. I am not oblivious to the critiques, some of which I consider quite well founded, but neither am I ignorant of the possible benefits these technologies offer. Chief among these, from my perspective, is the introduction of expanded notions of cognition that extend meaning-making practices beyond the human to the nonhuman and beyond the biological to the artificial. These are powerful resources to combat anthropocentrism, which has catalyzed human hubris and is a major factor in human practices that are destroying planetary ecosystems and endangering the futures of all living species, including humans.

In my view, the most significant ways in which ChatGPT and similar LLMs will relate to humans is through **intelligence augmentation**, amplifying human intelligence to achieve what would otherwise be impossible for human thinking and cognition. Many examples are already in evidence, such as GPT-4 being used to predict protein folding, an extremely complex problem that human thinking alone cannot solve; the results have been used to develop life-saving experimental drugs. [84] Especially now, with ChatGPT available free on the web, one need not be ultrarich or the CEO of a major company to benefit from its advice, nor does one need to have a world-critical problem to take advantage of it. For example, I was stumped on what to get my four-year-old granddaughter for her birthday, so I asked ChatGPT, and it instantly came up with a dozen good suggestions—four of which I actually used.

Can these programs be used for illegal, unethical, and even evil purposes? **Of course they can.** Most tech companies have tried to put guardrails around their programs to prevent the most obvious abuses (how to murder someone and get away with it, for example), but players big and small will no doubt find ways around them. As with every technology, it is a case of weighing the benefits and the costs, which include not only exploitation by bad actors but also the environmental damages of the enormous time and energy resources it takes to run these programs (**a point made in the "Stochastic Parrots" article**).

Used wisely, however, these programs have enormous potential. In my view, they are not only a game changer but also an evolutionary intervention of enormous importance to the human species. **Artificial intelligence is nothing less than a way to evolve life by means other than life**. In closing, I will risk a prediction. Short of environmental collapse or nuclear war, from now on, the trajectories of human and artificial intelligence will evolve together. For better or worse (perhaps for better and worse), the course of our futures and those of AI, our nonhuman symbionts, will run together.



## BIBLIOGRAPHY

Agüera y Arcas, Blaise. "Artificial Neural Networks Are Making Strides Toward Consciousness, According to Blaise Agüera y Arcas." *Economist*, September 2, 2022. <https://www.economist.com/by-invitation/2022/09/02/artificial-neural-networks-are-making-strides-towards-consciousness-according-to-blaise-aguera-y-arcas>.

———. "Do Large Language Models Understand Us?" *Daedalus* 151, no. 2 (Spring 2022): 183–97. [https://doi.org/10.1162/daed\\_a\\_01909](https://doi.org/10.1162/daed_a_01909).

Albert, Scott, Jihoon Jang, Shanaathanan Modchalingam, Bernard Marius 't Hart, Denise Henriques, Gonzalo Lerner, Valeria Della-Maggiore, Adrian M. Haith, John W. Krakauer, and Reza Shadmehr. "Competition between Parallel Sensorimotor Learning Systems." *eLife* 11 (2022): e65361. <https://doi.org/10.7554/eLife.65361>.

Amoore, Louise, Alexander Campolo, Benjamin Jacobson, and Ludovico Rella. "A World Model: On the Political Logics of Generative AI." *Political Geography* 113 (2024): 103134.

Bajohr, Hannes. "Artificial and Post-Artificial Texts: On Machine Learning and the Reading Expectations Towards Literary and Non-Literary Writing." BMCTT Working Papers No. 007. Department Arts, Media, Philosophy, University of Basel, March 2023. <https://hannesbajohr.de/wp-content/uploads/2023/03/Bajohr%2C%20Post-Artificial%20Writing.pdf>.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Schmittell. "On the Dangers of Stochastic Parrots: Can Language Models be Too Big?" In *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2021. <https://doi.org/10.1145/3442188.3445922>.

Blackiston, Douglas J., Emma Lederer, Sam Kriegman, Simon Garnier, Joshua Bongard, and Michael Levin. "A Cellular Platform for the Development of Synthetic Living Machines." *Science Robotics* 6, no. 52 (2021): eabf1571. <https://www.science.org/doi/10.1126/scirobotics.abf1571>.

Brenner, Eric D., Rainer Stahlberg, Stefano Mancuso, Jorge Vivanco, František Baluška, and Elizabeth Van Volkenburgh. "Plant Neurobiology: An Integrated View of Plant Signaling." *Trends in Plant Science* 11, no. 8 (2006): 413–19. <http://dx.doi.org/10.1016/j.tplants.2006.06.009>.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, et al. "Sparks of Artificial General Intelligence: Early Experiments with GPT-4." Preprint, arXiv, last revised April 13, 2023. <https://doi.org/10.48550/arXiv.2303.12712>.

Calvo, Paco. *Planta Sapiens: The New Science of Plant Intelligence*. With Natalie Lawrence. W. W. Norton, 2022.

Deacon, Terrence. *Incomplete Nature: How Mind Emerged from Matter*. W. W. Norton, 2013.

———. "Shannon-Boltzmann-Darwin: Redefining Information, Part 1." *Cognitive Semiotics* 1, no. S1 (Fall 2007): 123–48.

———. "Shannon-Boltzmann-Darwin: Redefining Information, Part 2." *Cognitive Semiotics* 2 (September 2008): 169–96.

———. *The Symbolic Species: The Co-evolution of Language and the Brain*. W. W. Norton, 1997.

Dhar, Payal. "GPT Protein Models Speak Fluent Biology: Deep-Learning Models Design Artificial Proteins for Tricky Chemical Reactions." *IEEE Spectrum*, February 1, 2023. <https://spectrum.ieee.org/ai-protein-design>.

Dresp-Langley, Britta. "Why the Brain Knows More than We Do: Non-Conscious Representations and Their Role in the Construction of Conscious Experience." *Brain Science* 2, no. 1 (2011): 1–21. <https://doi.org/10.3390/brainsci2010001>.

Fields, Chris, and Michael Levin. "How Do Living Systems Create Meaning?" *Philosophies* 5, no. 4 (2020): 36. <https://doi.org/10.3390/philosophies5040036>.

Gagliano, Monica, Michael Renton, Mariel Depczynski, and Stefano Mancuso. "Experience Teaches Plants to Learn Faster and Forget Slower in Environments Where it Matters." *Oecologia* 175 (2014): 63–72. <https://doi.org/10.1007/s00442-013-2873-7>.

Gagliano, Monica, Michael Renton, Nili Duvdevani, Matthew Timmins, and Stefano Mancuso. "Out of Sight But Not out of Mind: Alternative Means of Communication in Plants." *PLoS ONE* 7, no. 5 (2012): e37382. <https://doi.org/10.1371/journal.pone.0037382>.

Gagliano, Monica, Vladyslav V. Vyazovskiy, Alexander A. Borbély, Martial Depczynski, and Ben Radford. "Comment on 'Lack of Evidence for Associative Learning in Pea Plants.'" *eLife* 9 (2020): e61141. <https://elifesciences.org/articles/61141#x792b65f5>.

Gagliano, Monica, Vladyslav V. Vyazovskiy, Alexander A. Borbély, Mavra Gri-monprez and Martial Depczynski, "Learning by Association in Plants," *Scientific Reports* volume 6, Article number: 38427 (2016).

Grassini, Simone, Suvi K. Holm, Henry Railo, and Mika Koivisto. "Who Is Afraid of the Invisible Snake? Subjective Visual Awareness Modulates Posterior Brain Activity for Evolutionarily Threatening Stimuli." *Biological Psychology* 121, Part A (2016): 53–61. <https://doi.org/10.1016/j.biopsycho.2016.10.007>.

Haraway, Donna J. "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective." *Feminist Studies* 14, no. 3 (Autumn 1988): 575–99. <https://doi.org/10.2307/3178066>.

Hayles, N. Katherine. *Bacteria to AI: Human Futures with Our Nonhuman Symbionts*. University of Chicago Press, 2025.

———. *Unthought: The Power of the Cognitive Nonconscious*. University of Chicago Press, 2017.

Heaven, Will Douglas. "AI Hype Is Built on High Test Scores. Those Tests Are Flawed." *MIT Technology Review*, August 30, 2023. <https://www.technologyreview.com/2023/08/30/1078670/large-language-models-arent-people-lets-stop-testing-them-like-they-were/>.

Hoffmeyer, Jesper. *Biosemitotics: An Examination into the Signs of Life and the Life of Signs*. University of Chicago Press, 2008. Reprint, University of Scranton Press, 2009.

———. *Signs of Meaning in the Universe*. Indiana University Press, 1997.

Huddleston, Tom, Jr. "Bill Gates Watched ChatGPT Ace an AP Bio Exam and Went into a State of Shock." *CNBC*, August 11, 2023. <https://www.cnn.com/2023/08/11/bill-gates-went-in-a-state-of-shock-after-chatgpt-aced-ap-bio-exam.html>.

James, Henry. *The Figure in the Carpet*. FreeRiver Community, 2024. Originally published in 1896.

Lemoine, Blake. "Is LaMDA Sentient?—An Interview." *Medium*, June 11, 2022. <https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917>.

Lettvin, J. Y., Humberto R. Maturana, Warren S. McColluch, and Water H. Pitts. "What the Frog's Eye Tells the Frog's Brain." *Proceedings of the Institute for Radio Engineers* 47, no. 11 (November 1959): 1940–51.

Levin, Michael, and Daniel C. Dennett. "Cognition All the Way Down: Biology's Next Great Horizon Is to Understand Cells, Tissues and Organisms as Agents with Agendas (Even if Unthinking Ones)." *Aeon*, October 14, 2020. <https://aeon.co/essays/how-to-understand-cells-tissues-and-organisms-as-agents-with-agendas>.

Libet, Benjamin, and Stephen M. Kosslyn. *Mind Time: The Temporal Factor in Consciousness*. Harvard University Press, 2005.

Margulis, Lynn, and Dorion Sagan. *What is Life?* University of California Press, 2000.

Maturana, Humberto R. and Francisco J. Varela. 1980. *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht: D. Reidel.

Markel, Kasey. "Lack of Evidence for Associative Learning in Pea Plants." *eLife* 9 (June 2020): e57614. <https://elifesciences.org/articles/57614>.

Mitchell, Amir, Gal H. Romano, Bella Groisman, et al. "Adaptive Prediction of Environmental Changes by Microorganisms." *Nature* 460, no. 7252 (2009): 220–24. <https://doi.org/10.1038/nature08112>.

Murugan, Nirosha J., Daniel H. Kaltman, Paul H. Jin, et al. "Mechanosensation Mediates Long-Range Spatial Decision-Making in an Aneural Organism." *Advanced Materials* 33, no. 34 (2021): e2008161. <https://doi.org/10.1002/adma.202008161>.

Nagel, Thomas. "What Is It Like to Be a Bat?" *The Philosophical Review* 83, no. 4 (October 1974): 435–50.

Peirce, Charles Sanders. *Collected Papers*, vols. VII–VIII, edited by Arthur W. Burks. Harvard University Press, 1958.

Peirce, Charles Sanders. *The Essential Peirce: Selected Philosophical Writings*. Volume 1, edited by Nathan Houser and Christian Kloesel. Indiana University Press.

Powers, Richard. *The Overstory*. W. W. Norton, 2018.

Rouleau, Nicolas. “Comparative Cognition and the Multiple Realizability of Minds.” presented at the Other Minds workshop, Arizona State University, AZ, April 5, 2024.

Rouleau, Nicolas, and Michael Levin. “The Multiple Realizability of Sentience in Living Systems and Beyond.” *eNeuro* 10, no. 11 (2023): ENEURO.0375-23.2023. <https://doi.org/10.1523/ENEURO.0375-23.2023>.

Rouleau, Nicolas, Nirosha Murugan, and David Kaplan. “Towards Cognition in a Dish.” *Trends in Cognitive Science* 25, no. 4 (2021): 294–301. <https://doi.org/10.1016/j.tics.2021.01.005>.

Sacks, Oliver. *The Man Who Mistook His Wife for a Hat*. Vintage, 2021.

Simard, Suzanne W. “Mycorrhizal Networks Facilitate Tree Communication, Learning, and Memory.” In *Memory and Learning in Plants*, edited by Frantisek Baluska, Monica Gagliano, and Guenther Witzany. Springer Publishing, 2018.

Spivak, Gayatri C. “Can the Subaltern Speak?” *Die Philosophin* 14, no. 27 (1988): 42–58. <https://doi.org/10.5840/philosophin200314275>.

Stahlberg, Rainer. “Historical Overview on Plant Neurobiology.” *Plant Signaling and Behavior* 1, no 1 (January–February 2006): 6–8. <https://doi.org/10.4161/psb.1.1.2278>.

Stiegler, Bernard. *Technics and Time*. Vol. 1, *The Fault of Epimetheus*. Stanford University Press, 1998.

Van Le, Quan, Lynne A. Isbell, Jumpei Matsumoto, et al. “Pulvinar Neurons Reveal Neurobiological Evidence of Past Selection for Rapid Detection of Snakes.” *Proceedings of the National Academy of Sciences of the United States of America* 110, no. 47 (2013): 19000–19005. <https://doi.org/10.1073/pnas.1312648110>.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. “Attention Is All You Need.” In *Advances in Neural Information Processing Systems* 30 (NIPS), Long Beach, CA, 2017. [https://papers.nips.cc/paper\\_files/paper/2017/hash/3f5ee243547dee91fb-d053c1c4a845aa-Abstract.html](https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fb-d053c1c4a845aa-Abstract.html).

Verworn, Max. *Physiologisches Praktikum für Mediziner*, 2nd ed. G. Fisher, 1912.

Von Uexküll, Jacob. *A Foray into the Worlds of Animals and Humans: With a Theory of Meaning*. University of Minnesota Press, 2010.

Walker, Sara. “AI Is Life: Technology is Not Artificially Replacing Life—It Is Life.” *Noëma*, April 27, 2023. <https://www.noemamag.com/ai-is-life/>.

War, Abdul Rashid, Michael Gabriel Paulrai, Tariq Ahmad, et al. “Mechanisms of Plant Defense Against Insect Herbivores.” *Plant Signaling and Behavior* 7, no. 10 (2012): 1306–20. <https://doi.org/10.4161/psb.21663>.

Weatherby, Leif, and Brian Justie. “Indexical AI,” *Critical Inquiry* 48, no. 2 (2022): 381–415. <https://doi.org/10.1086/717312>.

Whorf, Benjamin. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. Edited by John B. Carroll. MIT Press, 1956.

Wolfram, Stephen. *What Is ChatGPT Doing . . . and Why Does It Work?* Barnes and Noble, 2023.

Yong, Ed. *An Immense World: How Animal Senses Reveal the Hidden Realms Around Us*. Random House, 2023.

Zhu, Xinwen, Emily R. Hager, Chuqiao Huan, and Allyson E. Sgro. “Leveraging the Model-Experiment Loop: Examples from Cellular Slime Mold Chemotaxis.” *Experimental Cell Research* 418, no. 1 (2022): 113218. <https://doi.org/10.1016/j.yexcr.2022.113218>.