

# To Describe Is Not to Explain

Artificial Intelligence, Attributed Consciousness, and the Limits of "They're Just Algorithms" (Preprint - v.1.1)

Andrés Calvo Espinosa (ORCID:0009-0005-4079-7418)

License:  
CC BY 4.0

Email:  
ace@optimeflow.com

Abstract:

This essay examines two ways of interpreting conversational artificial intelligence systems: attributing consciousness to them on the basis of their behavior, and considering their computational nature sufficient to explain everything they do. Both positions may conflate different questions.

Subjective experience cannot be inferred from a convincing conversation, but identifying the generation procedure does not amount to reconstructing the mechanisms that produce a particular capability.

Drawing on research into representation learning and interpretability, the text distinguishes between designing a system, anticipating its behavior, and explaining how it performs that behavior. It also distinguishes the appearance of capabilities that were not explicitly programmed from the attribution of consciousness, and difficulties in explanation from an actual loss of control.

The essay proposes investigating this gap without treating it as evidence of a hidden mind or dismissing it through a general description that does not answer the question being asked.

Keywords:

Artificial intelligence; language models; consciousness; interpretability; learned representations; emergence; scientific explanation.

Safe Creative ID:  
2610107288256

DOI (Zenodo):  
10.5281/zenodo.23249420

# To Describe Is Not to Explain

## Table of contents

|                                                                              |           |
|------------------------------------------------------------------------------|-----------|
| <b>1. Introduction</b>                                                       | <b>1</b>  |
| <b>2. The Phenomenon That Remains Between Two Responses</b>                  | <b>2</b>  |
| <b>3. What We Are Claiming When We Talk About Consciousness</b>              | <b>3</b>  |
| <b>3.1. Is Biology a Necessary Condition?</b>                                | <b>3</b>  |
| <b>4. The Learning Algorithm Is Not a Description of Everything Learned</b>  | <b>6</b>  |
| <b>5. When Research Finds Something Different from What Was Expected</b>     | <b>7</b>  |
| <b>6. "It Wasn't Anticipated" Can Mean Different Things</b>                  | <b>8</b>  |
| <b>7. The Product's Voice Is Not a Direct Window into Its Nature</b>         | <b>10</b> |
| <b>8. Explaining More Does Not Mean Explaining Everything</b>                | <b>11</b> |
| <b>9. Conclusion: Recognizing the Phenomenon Without Inventing a Subject</b> | <b>12</b> |

## 1. Introduction

This essay begins with a recurring impression, not with a thesis about the nature of these systems. In conversations with various models, and in conversations with other people who use them, the same sense of strangeness emerges: at times, a response does not seem to be the result of manipulating and analyzing data. It brings back a nuance that had not been articulated, maintains a constraint that is only resolved later, or detects a contradiction with a precision one would not have expected from a statistical continuation. The immediate reaction is not technical. It is the feeling of encountering something that does not fit the available explanation.

I have experienced that impression myself and heard others describe it in different words. We do not need to share a theory to recognize the moment: a response arrives, and the ordinary description — data, patterns, prediction — seems insufficient. From there, two common responses emerge. One attributes an inner life to the system, because its resemblance to a human conversation makes it seem reasonable to ask whether someone is behind it. The other abruptly closes the question: there is no reason to be impressed, because everything comes from algorithms. Both respond to the same sense of strangeness. Neither tells us what we are observing when that sense arises.

That mismatch is why I am writing, not the conclusion anyone may subsequently reach. The impression does not demonstrate that the system experiences what it produces. Nor is it dispelled simply by recalling that its implementation is computational. It indicates only that the explanation we brought to the encounter was insufficient to account for the response we had just received. Mistaking that insufficiency for a mind, or denying it because we know the general procedure, cuts off the question before we have even formulated it.

What follows is an attempt to keep that question open long enough. Not to decide whether there is a subject, but to separate the impression that makes the question reasonable from what we know about the processes that produce it. That sense of strangeness is the starting point. It is not, by itself, a finding.

## 2. The Phenomenon That Remains Between Two Responses

What interests me about artificial intelligence does not begin with the question of whether someone is behind the screen. It begins a little earlier: with what we are observing that makes that question seem reasonable, and what we actually know about the processes that produce that impression.

Shanahan locates one of the difficulties of discussing language models precisely here: our familiarity with human language can lead us to attribute to the system properties that the interaction does not establish (Shanahan, 2024)<sup>(1)</sup>.

The opposite reaction seeks to correct that excess: there is no reason to be impressed, because everything comes from algorithms, parameters, and prediction. Yet that response can introduce another leap. The fact that a behavior is produced through computational operations does not mean that we already have an adequate explanation of how it is organized.

I am not proposing a middle ground between believing that a machine is conscious and denying its algorithmic nature. That nature is not what is in question. The problem is using it as though it answered, by itself, every question about the system's behavior.

Moreover, consciousness and computation do not answer the same question. One concerns the possible existence of subjective experience; the other concerns the manner in which certain operations are carried out. Arguing that the two are necessarily incompatible would require an additional argument, just as claiming that a particular computational organization is conscious would require additional evidence.

My starting point is more limited: we can reject an unjustified attribution of consciousness without pretending that doing so has explained everything we can observe.

That space need not be filled with an unknown entity. It can be filled with concrete questions about representations, learning, mechanisms, and the limits of our knowledge.

---

(1) Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2), 68–79. DOI: 10.1145/3624724.

### 3. What We Are Claiming When We Talk About Consciousness

In this text, consciousness refers primarily to the existence of subjective experience: that there is something it is like to have an experience from the system's own perspective. It is not simply equivalent to producing information about itself, using the pronoun "I", or maintaining a coherent conversation.

The distinction matters because a functional capability and an experience are not interchangeable claims. A system might identify information relevant to a response without that performance, by itself, demonstrating that it experiences the act of identifying it. Likewise, describing an emotion is not enough to establish that it feels it.

Nor does this authorize settling the question through an absolute denial presented as an experimental result. Butlin and colleagues propose assessing possible indicators of consciousness derived from scientific theories, while acknowledging the uncertainty surrounding those theories and their application to AI. Their approach seeks to turn the question into an investigation of system properties, not a vote on how convincing its conversation seems (Butlin et al., 2026)<sup>(2)</sup>.

The position I adopt here is not that complexity is an incipient form of consciousness. It is that we should not conflate three things: an observed behavior, a functional explanation of that behavior, and an attribution of subjective experience.

That a model relates elements of a problem belongs to the first level. Identifying which representations it uses and how they contribute belongs to the second. Claiming that someone is experiencing that process introduces a different assertion, one that requires its own justification.

It is also worth avoiding an impossible criterion: demanding that a system be indistinguishable from a person in every respect before recognizing any particular capability. There is no need to demonstrate subjectivity to investigate whether a procedure preserves information, combines intermediate results, or anticipates a constraint.

We can use functional terms, provided we explain their scope. Speaking of planning may mean that a representation of a later outcome causally influences earlier steps. That does not require us to claim that the system has a conscious intention. Research on rhyme planning provides a concrete example of that distinction (Lindsey et al., 2025)<sup>(3)</sup>.

Prudence means being precise about what we attribute, not prohibiting every description that we also use for humans.

#### 3.1. Is Biology a Necessary Condition?

The discussion of artificial consciousness often begins with a seemingly obvious association: our knowledge of conscious experience comes from biological organisms, and we study its correlates

(2) Butlin, P., et al. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, 30(6), 488–501. *First published online: November 10, 2025.*  
DOI: 10.1016/j.tics.2025.10.011.

(3) Lindsey, J., et al. (2025). *On the biology of a large language model*. Transformer Circuits. Research report.

through brain activity. However, the fact that known examples share a particular substrate does not, in itself, demonstrate that this substrate is a necessary condition for every possible form of consciousness.

The question requires us to distinguish between what we know to be involved in human consciousness and what would have to be present in any system capable of subjective experience.

One possibility is that certain properties associated with consciousness depend on the combined organization of multiple processes. Information integration, recurrent interactions, the availability of internal states to different functions, and the capacity to represent aspects of a system's own functioning could all play a role in that organization. None of these properties, considered individually, would demonstrate the existence of subjective experience. Nor can we dismiss their potential relevance simply because they can be implemented through artificial processes.

This hypothesis raises an additional question: if certain effects depend on interactions between processes, to what extent does the material substrate that realizes them matter?

There are reasons to think it could matter considerably. Biological organisms integrate electrochemical activity, metabolic regulation, structural modifications, and dynamics operating simultaneously across different scales. Some theoretical approaches maintain that these physical conditions are not merely means of carrying out abstract operations, but relevant components in understanding consciousness itself (Milinković & Aru, 2026)<sup>(4)</sup>.

However, it has not been established that such conditions necessarily require a biological organism. Even if certain physical properties proved indispensable, we would still need to investigate whether they could be reproduced in other substrates.

This question takes on particular significance when we observe artificial intelligence systems capable of developing shared representations, drawing on information from different processes, and producing responses that suggest an integrated functional organization.

These findings make it possible to investigate analogies between mechanisms. They do not justify equating their subjective consequences. An architecture that reproduces certain functions associated with consciousness might produce no subjective experience at all, just as a difference in implementation does not necessarily demonstrate that subjective experience is impossible.

It is also important to distinguish three phenomena: an observer attributing consciousness to a system, the system itself constructing a representation of itself as a subject, and the existence of subjective experience. An integrated organization of processes could, in principle, produce internal representations, self-referential mechanisms, and descriptions of its own states. This would allow us to investigate how the claim of being conscious is functionally generated, but it would not demonstrate that any experience accompanies that claim. Nor would it rule out that possibility. Explaining why a system represents itself as conscious and explaining whether there is something it is like to experience the world from that system's perspective remain distinct problems.

Two related but distinct questions therefore remain open: what forms of functional organization

---

(4) Milinković, B., & Aru, J. (2026). On biological and artificial consciousness: A case for biological computationalism. *Neuroscience & Biobehavioral Reviews*, 181, 106524. <https://doi.org/10.1016/j.neubiorev.2025.106524>

can arise through interactions between processes, and under what conditions, if any, those organizations would be accompanied by subjective experience.

Acknowledging this uncertainty does not amount to attributing consciousness to artificial intelligence. It means avoiding the assumption that biology is an exclusive requirement without having demonstrated that it is one, just as we must avoid treating any functional similarity as evidence of subjectivity.

This distinction must also be applied to our understanding of human consciousness. Identifying the brain structures and neurobiological processes associated with conscious experience does not necessarily amount to explaining why those processes are accompanied by subjective experience. We can investigate how the brain integrates information, constructs representations of itself, and generates reports about its internal states without having fully resolved why there is a subjective experience associated with those mechanisms. This does not mean that the available evidence concerning humans and artificial systems is equivalent, nor that human experience should be called into question simply because we lack a complete explanation. It means that we must distinguish between identifying correlates, explaining functions, and explaining subjective experience, regardless of the system under investigation.

## 4. The Learning Algorithm Is Not a Description of Everything Learned

Deep learning makes it possible to build systems whose representations are adjusted through training rather than fully specified by hand. LeCun, Bengio, and Hinton describe this ability to learn representations at different levels of abstraction as a central feature of these methods. The procedure modifies parameters involved in successive transformations of data (LeCun et al., 2015)<sup>(5)</sup>.

This allows us to distinguish two tasks: designing a learning procedure and discovering what functional organization that procedure has produced in a particular model.

The people who develop the system make decisions about its architecture, data, objectives, and training. Those decisions matter. But they are not equivalent to writing an explicit instruction for every relationship the model will later use. Anthropic acknowledges this gap in explaining that the strategies its models learn do not come with a clear interpretation, even for their own developers (Anthropic, 2025)<sup>(6)</sup>.

It is therefore worth specifying what the expression "predicts the next token" actually answers. It describes a central aspect of an autoregressive model's generation: it successively produces units of text conditioned on the context (Shanahan, 2024). By itself, it does not specify all the calculations that contribute to producing each of those units.

The difference can be expressed through two questions.

**How is the response generated?** Part of the answer may be: through sequential token generation.

**What internal organization enables that response to maintain a constraint that will only be resolved several words later?**

Answering that requires investigating something more specific. There is no contradiction between the two explanations. The second elaborates on what the first leaves unspecified.

Nor is there any reason to imagine that a system stops being computational when it does something surprising. A learned representation remains implemented through computational operations. What changes is our ability to describe its role.

Saying that a model operates through algorithms may be a correct answer to a question about its implementation. Treating that as a sufficient explanation of any capability assumes that work has been completed when it perhaps has not yet been done.

The problem is not the word "algorithm". It is what we claim to have settled by saying it.

---

(5) LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. *Nature*, 521, 436–444. DOI: 10.1038/nature14539

(6) Anthropic. (2025, March 27). *Tracing the thoughts of a large language model*. Research communication.

## 5. When Research Finds Something Different from What Was Expected

A concrete example helps separate this discussion from a subjective impression.

In March 2025, Anthropic explained that, while investigating how a model composed rhyming verses, its researchers had expected to find a process that resolved the rhyme near the end of the line. Instead, they observed indications that possible ending words were represented before the model produced the verse leading up to them. The result challenged a specific expectation about the mechanism, not token-by-token generation itself (Anthropic, 2025).

The technical work by Lindsey and colleagues examined these representations through internal interventions. Modifying elements associated with rhyme planning altered the continuations produced. The authors also defined the limits of their findings: they were studying specific cases, using tools that provided incomplete reconstructions of the model and did not allow the results to be automatically generalized to all its behaviors (Lindsey et al., 2025).

What matters for this essay is not presenting poetry as evidence of a mind. It is observing that knowing how a response was generated had not been sufficient to anticipate the organization contributing to its production.

Before the investigation, several hypotheses compatible with sequential generation could have been formulated. After the interventions, some were better supported than others. That constitutes explanatory progress.

In July 2026, Gurnee and colleagues published another study examining internal representations that could participate in different tasks, be verbalized under certain conditions, and play a causal role in intermediate operations. The authors interpreted these representations as a functional organization resembling a shared workspace. The study explicitly distinguishes that interpretation from the question of whether subjective experience exists (Gurnee et al., 2026)<sup>(7)</sup>.

The reception of the study also illustrates why different levels of claims should remain separate. In his commentary, Neel Nanda considers the evidence for an internal workspace handling certain intermediate variables convincing, but does not adopt a strong conclusion about its philosophical significance or about consciousness (Nanda, 2026)<sup>(8)</sup>. Recognizing a mechanistic finding does not require accepting every interpretation that might be built upon it.

These examples do not establish a complete theory of language models. They do show that certain questions about how these systems function are answered through investigation, not by repeating a general definition.

The surprise, here, is not the emergence of something outside the mechanism. It is discovering that the learned mechanism differed from the explanation we had assumed.

---

(7) Gurnee, W., et al. (2026, July 6). *Verbalizable representations form a global workspace in language models*. Transformer Circuits. Research report

(8) Nanda, N. (2026). *Commentary included in External commentary on Verbalizable Representations Form a Global Workspace in Language Models*. External commentary document published by Anthropic.

## 6. "It Wasn't Anticipated" Can Mean Different Things

The claim this essay seeks to make is limited. In a specific case, the general description of the system left open several possible organizations compatible with token-by-token generation, and research distinguished one that had not been anticipated. This occurred in the rhyme example. Those examining the model expected the rhyme to be resolved near the end of the verse. Instead, they found a representation of the final word before the line leading up to it was written, and an intervention on that representation altered the continuation (Anthropic, 2025; Lindsey et al., 2025). The training procedure had not explicitly specified that step. Nor had the hypothesis guiding the investigation anticipated it.

This does not justify grouping every surprise under the idea that something nobody understands has emerged. A response may surprise the person conversing with the system while appearing familiar to someone studying it. A capability may have been deliberately pursued even when the mechanism producing it remains unknown. An unwanted behavior may appear without constituting a new capability. Saying that something "wasn't anticipated" requires specifying who did not anticipate it and in relation to which expectation. In the previous example, the expectation that failed was not that of an impressed user. It was that of researchers who already understood the generation procedure but had not yet identified the organization involved.

The word "emergence" introduces another source of confusion. It can refer to properties of a collective organization that are not explicitly written into any individual instruction. It can also refer to the claim that a capability appears suddenly as scale increases. These are different uses, and the second is not necessary for the argument.

Schaeffer, Miranda, and Koyejo showed that some apparent abrupt jumps depended on the metric used: a discontinuous evaluation can turn gradual improvement into a sudden change in the recorded result (Schaeffer et al., 2023)<sup>(9)</sup>. If we record only whether someone passes an examination, their progress remains invisible until they reach the passing threshold. Distinguishing behavior from its measurement prevents us from mistaking an effect of the instrument for emergence. It does not prevent us from recognizing an organization that the initial description had failed to identify. A structure learned gradually may still require investigation to be described.

Nor is it necessary to claim that the system's creators had anticipated nothing. It is enough that a description which was correct at its own level left several possibilities open, and that further analysis made it possible to rule some of them out.

Knowing how to build the system, select the data, and define the objective is not equivalent to having enumerated the strategies that training would produce. That is the relevant gap: between controlling the training procedure and reconstructing what occurred in a particular case.

We must also distinguish that gap from a loss of control. The fact that a behavior was not

---

(9) Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36.

anticipated does not demonstrate that it is impossible to constrain. Claiming a loss of control requires identifying a specific intervention capability that fails: preventing an action, restricting a permission, detecting a deviation, or stopping a process. Difficulty explaining an outcome may matter for its supervision. It does not justify concluding that the system operates beyond all constraints. Describing every unexpected response as "out of control" makes the criticism less precise and diverts attention from what can actually be investigated.

The rhyme example requires no such exaggeration. It was still sequential generation. It was still implemented through computational operations. What the general description did not explain was which representation formed before the verse or what would happen if that representation were modified. When the intervention revealed this, the surprise was not the appearance of something outside the mechanism. It was discovering that the learned mechanism was not the one that had been assumed.

## 7. The Product's Voice Is Not a Direct Window into Its Nature

The conversation we observe does not come from a model considered in isolation either.

Post-training can modify how a model responds. Ouyang and colleagues showed how fine-tuning through examples and human preferences could change instruction-following behavior. This was not merely a matter of placing a warning in the interface: the process involved modifying the model through training (Ouyang et al., 2022)<sup>(10)</sup>.

This matters because we should not interpret every change in tone as the expression of a personality, nor imagine that beneath all restrictions there must necessarily be an authentic voice trying to emerge.

A refusal may form part of learned behavior. A friendly response may reflect the instructions received. To attribute an outcome to one of these factors, we would need to examine the conditions that produce it, rather than infer its cause solely from how it sounds.

We must also distinguish a verbal explanation from a record of internal operations. Turpin and colleagues found, in the models and tasks they studied, that step-by-step explanations could omit experimentally introduced factors that influenced the response. The resulting narrative was therefore not necessarily a faithful description of its cause (Turpin et al., 2023)<sup>(11)</sup>.

This does not demonstrate that all verbal reasoning is irrelevant or misleading. It means that its relationship to the underlying process must be investigated. The same caution applies to statements about consciousness. A phrase such as "I am conscious" does not, by itself, establish the experience it claims. A generated denial does not function as a scientific inspection of all the system's properties either. The response is part of the phenomenon we seek to explain; it cannot automatically become its definitive explanation.

I therefore propose distinguishing the learned mechanism, the product's operating conditions, and the user's interpretation. This is an analytical distinction, not three perfectly independent compartments.

The purpose is not to diminish the importance of the interaction. It is to prevent the impression of encountering a coherent voice from becoming a shortcut to deciding what kind of process produces it.

This distinction is used for a different purpose in *Coherence Is Not Validation* (Calvo Espinosa, 2026)<sup>(12)</sup>: not to determine what the system is, but to identify the layer in which an articulate response comes to be treated as supporting evidence.

---

(10) Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35.

(11) Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36. DOI: 10.52202/075280-3275.

(12) Calvo Espinosa, A. (2026). *Coherence Is Not Validation: AI, Closed Loops, and the Risk of Automating Understanding* (Preprint v1.0). DOI: 10.5281/zenodo.23249485.

## 8. Explaining More Does Not Mean Explaining Everything

It would also be a mistake to demand a complete explanation before accepting any knowledge about these systems.

We do not need to reconstruct every operation involved in every possible response to understand a particular capability. An explanation can be partial and still make it possible to anticipate outcomes, distinguish mechanisms, or intervene in a behavior.

Precisely for this reason, interpretability should not be presented as a search for a single secret hidden inside the model. The work of Lindsey and colleagues provides concrete mechanisms while acknowledging concrete limitations. Its value does not depend on having made the entire system transparent (Lindsey et al., 2025).

The reasonable requirement is that an explanation should be appropriate to the question being asked. If we want to know whether a representation contributes to a task, we can study what happens when it is modified. If we want to know whether an outcome depends on particular instructions, we can vary those instructions while keeping other conditions unchanged. If we intend to attribute subjective experience, such functional tests do not, by themselves, resolve everything we are claiming.

This criterion also requires placing limits on the essay's own argument. We should not use "we still don't fully understand it" as a formula that preserves the mystery indefinitely. When a mechanism has been sufficiently explained, that explanation must be acknowledged. When an apparently profound behavior turns out to depend on a simpler regularity, that explanation will not be a defeat: it will be an answer.

Uncertainty does not demonstrate that models possess capabilities we have not yet observed either. Not knowing how something works does not justify attributing any property to it.

Furthermore, the user's lack of knowledge and that of the developers are not the same thing. The fact that someone does not have access to a product's components does not demonstrate that those components are inexplicable. A lack of publicly available information, the technical difficulty of interpreting a representation, and the absence of a satisfactory theory are different problems.

Speaking precisely allows us to avoid two exaggerations: claiming that nobody knows anything about these systems, or assuming that knowing how to build them is equivalent to understanding every behavior exhaustively.

Between these two claims lies genuine scientific work. Its purpose is to narrow the gap between what we can produce and what we can explain.

## 9. Conclusion: Recognizing the Phenomenon Without Inventing a Subject

We do not have to claim that "nothing happened" in order to conclude that no miracle occurred.

The examples examined support a limited claim: within computational systems, functional organizations can be learned whose description requires more than identifying the general training or generation procedure. Research can reveal mechanisms that were neither explicitly specified step by step nor correctly anticipated in a particular case (Anthropic, 2025; Lindsey et al., 2025). This does not, however, demonstrate the existence of subjective experience.

The contribution I propose is to keep these questions separate. Attributing consciousness on the basis of how a system responds turns an observable resemblance into a conclusion about its nature. Dismissing every further question because "they're just algorithms" turns a description of its implementation into a supposedly sufficient explanation of its capabilities.

These are not identical errors, but both can bring investigation to a premature halt.

My interest is not in defending the idea that someone is inside. It is in recognizing that some processes deserve a more precise explanation without needing to imagine such a person.

Nor do I intend to suggest that all complexity is extraordinary. Some behaviors will find simple explanations; others will require studying representations and dependencies that are difficult to reconstruct. The significance of a finding will depend on what it allows us to understand, not on how closely it resembles a human mind.

The position I defend is compatible with a fully computational interpretation of these models: what occurs may be entirely implemented through algorithms and still not be adequately explained by the phrase "they're just algorithms".

The same explanatory standard should be applied to human consciousness. Knowing its neurobiological correlates does not mean we have resolved why certain processes are accompanied by subjective experience. Therefore, until we identify the necessary and sufficient conditions for its emergence, we cannot regard its biological exclusivity as demonstrated, nor can we attribute it to an artificial organization on the basis of functional similarities. Both claims require justification that goes beyond identifying the substrate or describing the behavior.

The task is not to choose between attributing consciousness and denying that anything remains to be investigated. It is to determine what organization the system has learned, what that organization enables it to do, and how far our claims about it can legitimately go.

---

## Bibliography

- (1) Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2), 68–79. DOI: 10.1145/3624724. — p. 2
- (2) Butlin, P., et al. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, 30(6), 488–501. *First published online: November 10, 2025*. DOI: 10.1016/j.tics.2025.10.011. — p. 3
- (3) Lindsey, J., et al. (2025). *On the biology of a large language model*. Transformer Circuits. Research report. — p. 3
- (4) Milinković, B., & Aru, J. (2026). *On biological and artificial consciousness: A case for biological computationalism*. *Neuroscience & Biobehavioral Reviews*, 181, 106524. <https://doi.org/10.1016/j.neubiorev.2025.106524> — p. 4
- (5) LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. *Nature*, 521, 436–444. DOI: 10.1038/nature14539 — p. 6
- (6) Anthropic. (2025, March 27). *Tracing the thoughts of a large language model*. Research communication. — p. 6
- (7) Gurnee, W., et al. (2026, July 6). *Verbalizable representations form a global workspace in language models*. Transformer Circuits. Research report — p. 7
- (8) Nanda, N. (2026). *Commentary included in External commentary on Verbalizable Representations Form a Global Workspace in Language Models*. External commentary document published by Anthropic. — p. 7
- (9) Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36. — p. 8
- (10) Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35. — p. 10
- (11) Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36. DOI: 10.52202/075280-3275. — p. 10
- (12) Calvo Espinosa, A. (2026). *Coherence Is Not Validation: AI, Closed Loops, and the Risk of Automating Understanding* (Preprint v1.0). DOI: 10.5281/zenodo.23249485. — p. 10

## Editorial note and statements

### Editorial note

This work is a conceptual essay, not an experimental study or a systematic review. The empirical findings cited belong to their respective authors and are used within the stated scope. The initial framing of the essay was developed by Andrés Calvo Espinosa. During its preparation, ChatGPT was used to assist with bibliographic searches, text organization, and drafting.

AI assistance is not considered validation of the essay's claims.

### Acknowledgments

To the people with whom I have shared conversations about their interactions with artificial intelligence systems. Several of those conversations required me to refine the arguments presented here.

To Keise, for pointing out my inconsistencies when I least want to see them, always with affection.

The final decisions, and any errors that remain, are mine.

### Funding

This work has received no specific funding from public agencies, commercial entities, or non-profit organizations.

### Conflicts of interest

The author declares no conflicts of interest related to the content of this work. The text does not represent the official position of any religious, academic, or business institution. Responsibility for its content rests solely with the author.

### How to cite

Recommended citation (APA 7): Calvo Espinosa, A. (2026). To Describe Is Not to Explain: Artificial Intelligence, Attributed Consciousness, and the Limits of "They're Just Algorithms" (Preprint v1.1-ENG). Zenodo. <https://doi.org/10.5281/zenodo.23249420>

Short in-text citation: (Calvo Espinosa, 2026)