

Coherence Is Not Validation

AI, Closed Loops, and the Risk of Automating Understanding

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

License:
CC BY 4.0

Email:
ace@optimeflow.com

Abstract:

This essay examines the difference between internal coherence and external validation in the use of generative artificial intelligence systems. It draws on observations of interactions in which the progressive reformulation of an interpretation may be perceived as confirmation, even when no independent evidence has been introduced.

The analysis distinguishes between the generation processes of models, their configuration as products, and the interpretation made by those who use them. It also examines how certain feedback loops can consolidate explanations that reinforce themselves internally without being sufficiently exposed to contradiction.

Through examples drawn from human–AI interaction, organizational interpretation, and operational documentation, the text proposes that automating a task may eliminate part of its cognitive function when that function consists precisely in formulating, testing, and sharing knowledge.

As a cross-cutting criterion, the essay proposes identifying which observations or outcomes would require an explanation to be revised before accepting its coherence as evidence of validity.

Keywords:

Generative artificial intelligence; external validation; internal coherence; confirmation bias; cognitive automation; critical thinking; organizational documentation; human–AI interaction.

Safe Creative ID:
2610107287822

DOI (Zenodo):
10.5281/zenodo.23249485

Coherence Is Not Validation

Table of contents

1. Introduction	1
2. Describing the System Is Not Explaining It Completely	2
3. Internal Coherence, External Validation, and Closed Feedback Loops	3
4. Automating Documentation Is Not the Same as Understanding	4
5. Three Layers for Interpreting a Response	5
6. The Shared Loop Between User and Model	6
7. When Critical Testing Works	7
8. Criteria for Testing and Revision	8
9. Conclusion: Recognizing When to Withhold Belief	9

1. Introduction

This text arises from repeated use, not from a preconceived thesis. Over the years, I have questioned several systems at different times and in different states of mind. On some occasions, I let myself be carried away by the language and emotion of the exchange. I knew that some of the things I was asserting to the machine were wrong. Even so, the response gave me back a more articulate version of what I had just said, and that articulation felt like agreement. The agreement was mine. The system did not need to tell me I was right: it only needed to continue within my framework, refine it, and return it to me.

That proves nothing about the nature of the model. It raises a different question: what happens to the reasoning of someone using the system when generation and assent remain enclosed within the same conversation? What follows seeks to separate those layers. It does not aim to settle the question. It aims to prevent the fluency of the exchange from taking the place of verification.

The proposed contribution is an applied conceptual synthesis: bringing these problems together to distinguish when a conversation provides a genuine test and when it merely develops an interpretation, and when automating a record preserves or displaces the activity of understanding that justified producing it. The experiences and examples serve to formulate these questions, not to establish how frequently the phenomenon occurs or to demonstrate its general effects.

2. Describing the System Is Not Explaining It Completely

Artificial intelligence companies do not merely market a system. They also market a description. "Model", "tool", "assistant", or "next-token prediction" are valid descriptions at certain levels of analysis. However, naming an operation or an assigned role is not the same as explaining everything we observe when we use the system.

The point is not to claim that a neural network and a brain are the same thing. They are not. Different processes can perform comparable functions without sharing the same mechanisms. Maintaining coherence, integrating information, or producing a response conditioned by previous context are functional descriptions; by themselves, they do not establish an identity between the systems that perform them.

This raises a semantic problem. If we use verbs such as remember, anticipate, correct, or reason when discussing these systems, we need to specify what we mean by them. Caution about attributing excessive agency does not require us to minimize observable capabilities. Nor does acknowledging those capabilities require us to attribute subjective experience to the system.

"Predicts the next token" identifies a central operation of these models. But it does not, by itself, specify the calculations and representations that produce a particular continuation. To examine their behavior, we must also ask what relationships the system has learned, how it uses them, under what conditions it fails, and which aspects of what we observe arise from training, instructions, tools, or the interface. A technically correct description may be insufficient for a question different from the one it answers.

An alternative description must, in turn, explain what it adds and which observations or arguments would require it to be revised. Our difficulty in reconstructing a behavior does not demonstrate an extraordinary property of the system: it may reflect a limitation of our knowledge.

Fluency does not demonstrate subjectivity. Shanahan warns that interacting with a conversational agent can create the illusion of engaging with a thinking creature, even though the system is not, by its nature, like us (Shanahan, 2024)⁽¹⁾. The similarity of the interaction is not sufficient to determine which mechanisms produce it or which properties can be attributed to the system.

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

3. Internal Coherence, External Validation, and Closed Feedback Loops

A system can generate a response, evaluate it through another layer of the same system, and produce a new response based on that evaluation. Repeating this process may improve its coherence. But coherence is not validation. When an explanation claims to describe the world, it requires a test that is not merely another reformulation of the same content: an observation, an experiment, a relevant source, or a verifiable consequence.

In this essay, "validation" does not mean a definitive guarantee of truth, but a check that is relevant and sufficiently independent for the type of claim being examined. The relevant independence does not consist of another person or application repeating the conclusion. It consists of ensuring that the criterion and the evidence do not depend solely on accepting the explanation being tested.

This is why the problem is not exactly that AI "validates itself". It is more accurate to say that certain uses can create closed feedback loops in which generation, evaluation, and correction reinforce one another without sufficient testing. Under these conditions, consistency across responses may be perceived as confirmation even though the evidential support for the explanation has not increased.

Internal review is not without value. It can uncover errors, rule out incompatible conclusions, or make the consequences of the premises explicit. The absence of new information does not make that work useless. What cannot be done is to treat its result, by itself, as proof that the premises correctly described the world.

The problem is not exclusive to artificial intelligence either. Nickerson examines how the search for and interpretation of evidence can favor pre-existing beliefs (Nickerson, 1998)⁽²⁾. An organization may engage in a similar process when it interprets any request for clarification as confirmation that someone is "obstructing the work". If different responses end up preserving the same judgment, it is worth asking what information would make it possible to revise it.

The analogy does not attribute the same mechanism, intention, or responsibility to an organization and a model. It identifies a possible structure of the process: the interpretation becomes protected from what should be capable of correcting it. Identifying that similarity is not sufficient to explain its causes in each case.

(2) Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220.
<https://doi.org/10.1037/1089-2680.2.2.175>

4. Automating Documentation Is Not the Same as Understanding

The same shift occurs when the function of a tool extends beyond the file it produces. I built an operational workspace, including a project management tool, for an internal team. The purpose was not merely to have an accurate written account of the work. It was to require each person to articulate what the project was, what was missing, and what depended on them. That effort was the mechanism I wanted to preserve. The record was the by-product. Writing was meant to help build an overall understanding of the whole.

The next request was to automate the transfer of information into the workspace. To have a system write the entries and leave the workspace populated. The time saved can be real. Cognitive offloading makes it possible to reduce the demands of a task through external actions and aids (Risko & Gilbert, 2016)⁽³⁾. The question, in this case, is how much of the original function is preserved by doing so. If the system drafts and records entries without effective review by the team, the workspace may be complete without participants ever having to articulate what they know and what they do not know. The team can point to the workspace and say that the project is documented. What cannot be inferred from that record alone is that a shared understanding exists.

Not every form of automation poses the same problem. A status already recorded in another tool — a date, a closure, an already-verified change — can be transferred without repeating the work behind it. The issue changes when automation replaces the step intended to have someone explain, in their own words, what is happening and what they do not know. If that effort is not preserved elsewhere in the process, the workspace may still appear complete while having lost part of its function as a tool for shared knowledge.

The risk becomes apparent when someone consults the workspace and mistakes the continuity of its prose for an understanding that the team has already developed and shared. The text may be complete. That does not establish that the team understands it. And if the system not only drafts the entry but also writes it to the record and marks it as closed without effective review, the unread paragraph becomes the status of the project. The fluency of the automated transfer takes the place of verification.

(3) Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>

5. Three Layers for Interpreting a Response

I develop the distinction between learned mechanisms, product configuration, and user interpretation elsewhere and do not repeat it here (Calvo Espinosa, 2026)⁽⁴⁾. One consequence is enough for the present discussion. In a conversation, these layers cannot be isolated by examining the tone: attributing an outcome to training, an instruction, a tool, or an expectation may require evidence that the exchange does not provide.

For this essay, it is the third layer that matters. The user sees a well-articulated response and decides what counts as evidence, correction, or confirmation. The loop described in the next section begins there: not when the model produces a continuation, but when that continuation is taken as support.

(4) Calvo Espinosa, A. (2026). *To Describe Is Not To Explain: Artificial Intelligence, Attributed Consciousness, and the Limits of “They’re Just Algorithms”* (Preprint v1.0). Zenodo. DOI: 10.5281/zenodo.23249420

6. The Shared Loop Between User and Model

One possible loop works as follows: the user provides an intuition, the model develops it, and the user recognizes that intuition in a more articulate form. The user takes that articulation as confirmation. The model receives the user's acceptance as part of the context and continues elaborating in the same direction. Eventually, successive turns appear to provide evidence, even though they may have merely multiplied the formulations of an initial premise. No intention to deceive is required for this shift to occur.

There is a related, though not identical, empirical precedent: Sharma and colleagues study forms of sycophancy in which assistants adapt their responses to users' opinions or reproduce their errors, and find indications that human preferences used in training may encourage such behavior (Sharma et al., 2024)⁽⁵⁾. This essay also focuses on the user's interpretation: the possibility that a continuation within an existing framework is received as though it provided evidence. The two phenomena may overlap; the assistant does not need to explicitly state that the user is right.

A composite example drawn from different experiences illustrates this process. A user writes that a colleague is "blocking the project" because they requested documentation before a meeting. The system organizes the account, assigns causes, and returns a version in which the request has already become a pattern of obstruction. The user recognizes their irritation expressed more clearly and takes that articulation as confirmation. In the next turn, the system receives that acceptance as context and proposes how to document the obstruction. At no point has anyone checked whether the request was a requirement, a delay, or something else. The relevant test would be to retrieve the original message and determine whether an applicable rule existed, without assuming in advance that the obstruction interpretation was valid.

This is why simply asking the model to "be critical" is not enough. An objection generated by the same system may be useful, but its appearance of independence does not guarantee genuine independence. Nor is adding citations sufficient: it is necessary to verify that the sources exist, that they support the claims being made, and that they have not been selected solely because they preserve the conclusion.

(5) Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askill, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards understanding sycophancy in language models. *International Conference on Learning Representations (ICLR)*. https://proceedings.iclr.cc/paper_files/paper/2024/hash/0105f7972202c1d4fb817da9f21a9663-Abstract-Conference.html

7. When Critical Testing Works

The loop described above does not characterize every interaction with a model. The presence of a tool, a source, or human review is not, by itself, sufficient to determine whether adequate testing has taken place. The difference lies in what is checked and whether the result can change the initial interpretation. The following scenarios are hypothetical: they serve to clarify the argument, not to present findings from a study.

Consider a variation of the previous example. The system retrieves the original message and the applicable rule. The documentation shows that requesting those materials was a prerequisite, not an additional obstacle introduced by the colleague. If the response withdraws that attribution because the evidence examined does not support it, the conversation has not merely reformulated the premise: it has made it possible to correct it. The test may have been performed within the same interface and with automated assistance. It would not be appropriate to describe that episode as confirmation of a closed loop.

Human review can also perform this function. Suppose a reviewer examines the documents, considers an alternative interpretation, and rejects the system's recommendation because it does not correspond to what happened. The reviewer's intervention does not consist of accepting the framework with a signature: it introduces an evaluation capable of changing the outcome. The fact that a review is carried out by a person does not guarantee this effect, but neither does it justify dismissing it.

Formal verification introduces an additional distinction. If a generated proof is correctly verified according to explicit rules and assumptions, its acceptance does not depend on whether a second account sounds convincing. It can be established that a conclusion follows within that framework. This does not demonstrate that the assumptions describe an empirical situation, but it does provide justification for the formal claim being examined. The fact that a verification has a limited scope does not make it circular; it prevents that verification from being used to certify something other than what it has actually checked.

Something similar may happen with the operational workspace. Suppose the system prepares entries, identifies uncertainties and contradictions, and the team reviews the decisions and explains their dependencies before approving them. If that procedure preserves or improves understanding, manual drafting was not indispensable. Automation would have changed the way the activity was carried out, not necessarily eliminated its function.

These cases define the limits of the warning. It is not enough to observe that a response was generated automatically, reviewed by the same system, or accepted by a person. We must examine what supporting evidence was incorporated, which errors were detected, and which conclusions changed. The question is not only how an interpretation can become closed to revision, but also which procedures make it possible to correct it and how we can recognize that such correction has occurred.

8. Criteria for Testing and Revision

The criterion for testing must be made explicit before a conclusion is considered validated, rather than adjusted afterward solely to protect it. This does not require retaining inadequate criteria: it requires justifying changes to them without making those changes depend exclusively on the conclusion one wishes to preserve.

This requirement is consistent with the importance Popper places on subjecting empirical theories to tests capable of contradicting them, rather than supporting them solely through compatible cases (Popper, 2002)⁽⁶⁾. In this essay, it is applied as a procedural rule, without requiring every conceptual proposal to be resolved through a single experiment.

For a claim about the world, it would be necessary to identify what observation would be incompatible with it, what predicted consequence might fail to occur, or what check would show that a source does not support what is attributed to it. For a conceptual interpretation, counterexamples and objections that require its distinctions to be revised also matter. A misused reference weakens the support for a claim; it does not automatically demonstrate that the claim is false.

In the operational workspace example, it would be necessary to examine more than the quality of the record or the time saved: whether participants can explain decisions, identify dependencies, and detect errors. If, under comparable conditions, automation preserved or improved those capabilities without requiring manual drafting, the idea that this step is necessary for understanding in that context would have to be revised. The function might have been preserved through another means.

To study the role of articulation, one possible test would be to compare responses containing the same claims, reasons, and sources, but differing in their degree of elaboration. It would be necessary to observe not only whether the user's confidence changes, but also whether they understand the argument better and whether their confidence is aligned with the correctness and evidential support of the claim. A clearer explanation could help someone understand a valid justification; increased confidence would not, by itself, constitute an error.

If a sufficiently informative comparison showed that greater elaboration does not reinforce unjustified confidence, the role attributed to that factor would have to be limited under the conditions studied. If it promoted more accurate understanding and better-founded confidence, that result should be recognized as a benefit, not reinterpreted to preserve the warning unchanged.

These proposals are neither results nor complete experimental protocols. They identify which parts of the argument require further research and what findings might delimit them. The criterion is not that every verification must come from a person, but that an explanation must not be allowed to determine, at its convenience, what counts as evidence in its favor.

(6) Popper, K. R. (2002). *The logic of scientific discovery*. Routledge. (Originally published in German as *Logik der Forschung* in 1934, with a 1935 imprint date; first English edition, prepared by the author, published in 1959). <https://www.routledge.com/The-Logic-of-Scientific-Discovery/Popper/p/book/9780415278447>

9. Conclusion: Recognizing When to Withhold Belief

Progress does not consist solely of producing more convincing responses. It also consists of developing better ways to recognize when we should not believe them and when we do have reasons to trust them.

The central distinction of this essay does not set artificial intelligence against human thought. It separates the development of an explanation from the supporting evidence that justifies accepting it, and the production of a document from the understanding that the document alone cannot establish. In both cases, automated assistance may contribute to the task or displace part of its function. Determining what has happened requires examining the process, not inferring it from who drafted the result.

The text offers a synthesis and a set of research questions, not a demonstration of the frequency or magnitude of these risks. Its usefulness will depend on whether it helps identify both the processes that protect a premise from revision and those that succeed in correcting it. Finding the latter should limit the application of the warning, not become another way of confirming it.

Coherence can improve an explanation. It does not exempt that explanation from being accountable to what it claims to explain.

Bibliography

- (1) Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2), 68–79.
<https://doi.org/10.1145/3624724> — p. 2
- (2) Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220. <https://doi.org/10.1037/1089-2680.2.2.175> — p. 3
- (3) Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688.
<https://doi.org/10.1016/j.tics.2016.07.002> — p. 4
- (4) Calvo Espinosa, A. (2026). *To Describe Is Not To Explain: Artificial Intelligence, Attributed Consciousness, and the Limits of “They’re Just Algorithms”* (Preprint v1.0). Zenodo. DOI: 10.5281/zenodo.23249420 — p. 5
- (5) Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Asbell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards understanding sycophancy in language models. *International Conference on Learning Representations* (ICLR).
https://proceedings.iclr.cc/paper_files/paper/2024/hash/0105f7972202c1d4fb817da9f21a9663-Abstract-Conference.html — p. 6
- (6) Popper, K. R. (2002). *The logic of scientific discovery*. Routledge. (Originally published in German as *Logik der Forschung* in 1934, with a 1935 imprint date; first English edition, prepared by the author, published in 1959).
<https://www.routledge.com/The-Logic-of-Scientific-Discovery/Popper/p/book/9780415278447> — p. 8

Editorial note and statements

Editorial note

This text is an essay of conceptual and exploratory analysis, focused on interactions with conversational systems based on language models. The experiences described, the composite example, and the hypothetical scenarios are used to formulate questions and criteria for revision; they do not constitute a systematic empirical study or allow results to be generalized to all systems or contexts of use.

During the preparation of this work, ChatGPT was used to assist with drafting, organizing the text, reviewing arguments, and proposing bibliographic references. The initial framing and the experiences described are the author's own. The author retains editorial control and assumes responsibility for the published version. 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). Coherence Is Not Validation: AI, Closed Loops, and the Risk of Automating Understanding (Preprint v1.0-ENG). DOI: 10.5281/zenodo.23249485

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