2 Versions

Beyond World Models: Rethinking Understanding in AI Models

(/pdf?id=s4nioVFfnj)

Tarun Gupta (/profile?id=~Tarun_Gupta4), Danish Pruthi (/profile?id=~Danish_Pruthi1)

Published: 08 Nov 2025, Last Modified: 05 Mar 2026 AAAI-26 Poster
Conference, Area Chairs, Senior Program Committee, Program Committee, Publication Chairs, Authors
Revisions (/revisions?id=s4nioVFfnj) BibTeX

Serve As Reviewer: Danish Pruthi (/profile?id=~Danish_Pruthi1)
Keywords: world-models, understanding, epistemology
Primary Keyword: PEAI: AI & Epistemology
TL;DR: We argue that world models are an inadequate lens for claiming that AI systems understand in a human-like way, using case studies from prior philosophical work to demonstrate this limitation.
Secondary Keywords: PEAI: Philosophical Foundations of AI, PEAI: Artificial General Intelligence
Abstract:
The AI community has shown substantial interest in the concept of world models: internal representations that simulate aspects of the external world, track entities and states, capture causal relationships, and enable prediction of consequences. This contrasts with representations based solely on statistical correlations. A key motivation behind this research direction is the argument that humans possess such mental world models, and finding evidence of similar representations in AI models might indicate that these models truly "understand" the world in a human-like way. In this paper, we use problems and case studies from the philosophy of science literature to critically examine whether the world model framework adequately characterizes human-level understanding. We focus on specific philosophical analyses where the distinction between world model capabilities and human understanding is most pronounced. While these represent particular views of understanding rather than universal definitions, they illuminate some important limitations in using world models as a lens to claim that AI models understand in a human-like way. By highlighting these distinctions, we hope to stimulate deeper discussion about the nature of understanding in both human and artificial contexts.

Country Of Institutions: India
Profile Policy Agreement: I confirm that all authors have up-to-date OpenReview profiles, including their current position, institution-affliated email address, and DBLP URL. I understand that submissions with incomplete author profiles will be subject to desk rejection.
Submission Number: 16917

Add: **Withdrawal** **Ethics Chair Author Comment**

Paper Decision

Decision by Program Chairs 08 Nov 2025, 00:31 (modified: 08 Nov 2025, 04:50)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=miSnaDHi8T)

Decision: Poster

Add: **Ethics Chair Author Comment**

Phase 2 AC Recommendation by Area Chairs

Phase 2 AC Recommendation by Area Chairs 29 Oct 2025, 15:09 (modified: 08 Nov 2025, 03:58)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=qtNAg2vp3i)

Metareview:

This is a fascinating case. The paper is primarily philosophical, and none of the reviewers are philosophers. However, very few of the AAAI attendees are philosophers. (Are the authors aware of the cost of attending AAAI? It is far higher than that of a philosophy conference, I think. Here are the fees for current AAAI members; membership itself is nontrivial: https://aaai.org/conference/aaai/aaai-26/registration/#member-tech (https://aaai.org/conference/aaai/aaai-26/registration/#member-tech); my apologies if you are already aware of this.)

The argument that "world knowledge" is different from "human level understanding" is one I would wish all AAAI attendees to hear. However, it is unclear that they would in fact hear it, rather than being hung up on the same objections the reviewers raised ("no experiments!" and "no crisp definitions"), because those are standards of the field.

Because this seems so important to me, and because much positive was said in the reviews, I recommend a compromise, namely a poster.

Acceptance Recommendation: This paper is in the bottom 25% of papers presented at a top tier venue like AAAI. (Weak accept recommendation.)

Confidence: 4: The AC is confident but not absolutely certain

Add: **Ethics Chair Author Comment**

Phase 2 SPC Recommendation by Senior Program Committee VmEz

Phase 2 SPC Recommendation by Senior Program Committee VmEz

25 Oct 2025, 04:56 (modified: 08 Nov 2025, 05:44)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee, Authors

Revisions (/revisions?id=MrogOgBcGc)

Metareview:

Important note from the SPC member writing this metareview: I believe that the use of LLM technology, which has extensively documented limitations, for a serious endeavor like evaluating submissions to one of the premier conferences in a discipline to be highly questionable. I cannot in good conscience make use of the material posted as “AI Reviews”, and have therefore elaborated all of my metareviews without taking it into consideration.

This submission received five reviews, three of which are low confidence, but the other two are not. All reviewers clearly did their best to evaluate the work, and all agree on a slightly negative overall evaluation. Though a subset of the reviewers based their negative views mostly on the fact that the paper is “only philosophical”, during Phase 2 the new reviews focused on delimitation issues that lead to a lack of solid footing for the paper’s arguments. Though the paper is clearly thought-provoking, it is not ready for publication in a venue like AAAI in its current form. Unfortunately, there was almost no discussion after the rebuttal phase.

Acceptance Recommendation: This paper is slightly below papers presented at a top tier venue like AAAI. (Weak reject recommendation.)
Confidence: 4: The SPC is confident but not absolutely certain

Add: **Ethics Chair Author Comment**

Rebuttal by Authors

Rebuttal by Authors (Tarun Gupta (/profile?id=~Tarun_Gupta4), Danish Pruthi (/profile?id=~Danish_Pruthi1))
10 Oct 2025, 18:25 (modified: 12 Oct 2025, 13:26)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Authors
Revisions (/revisions?id=x8sOzzStVD)

Rebuttal:
We thank the reviewers for their feedback and are glad they found our paper thought-provoking (Qnex, vjvP), insightful (vjvP, 13rf), highly persuasive and excellent (vjvP), well written (13rf), enlightening (v3ym) and **significant for the community** (v3ym). We address the main concerns below:

No empirical experiments (QHWY, L8Kx, 13rf): Our paper offers a critique of the world-model research direction using philosophical case studies. Such arguments do not hinge on, nor map easily to, empirical tests. Historically, AI progress has often stemmed from philosophical inquiry, e.g., the *Turing Test* and Searle's *Chinese Room*. Our work, submitted to the *Philosophy and Ethics of AI* track, aims to advance philosophical understanding, which does not require empirical experiments.

Lack of definitions (QHWY, Qnex, v3ym): Our paper deliberately avoids defining "understanding." We argue that providing a definition, even a so-called "operational definition," is strictly counter-productive. In epistemology, *understanding* is a contested concept with no agreed definition, and forcing one only produces false precision. Any operational definition would merely shift the problem to the defining terms, leading to an infinite regress unless we admit so-called "primitive" terms, that is, undefined terms [1]. See [1, 2] for how false precision in unsettled concepts can stifle philosophical progress. Hence, rather than impose an arbitrary definition on a concept still unsettled in the broader literature, we take an orthogonal approach—using case studies where aspects of understanding are apparent from prior analyses and showing how the world-model framework fails to capture them.

"Why would allowing for a concept of prime to be represented in a world model risk the approach's falsifiability?" (Qnex): Allowing ad infinitum changes in abstraction makes the notion of world models vague, and a vague theory cannot be falsified. If abstraction levels are unconstrained, any phenomenon can be retrofitted as a world model. Hence, we fix the abstraction level—consistent with prior work in the domains of our three case studies—and show that it fails to capture understanding.

Lastly, we note that reviewer **vjvP's glowing review** appears inconsistent with the score of 5 (marginally below threshold). We hope this was not an oversight.

[1] K. R. Popper, The Open Society and Its Enemies, Vol. II, Chap. 11, Sec. II.

[2] F. P. Ramsey, Philosophical Papers, Chap. 1.

Add: **Ethics Chair Author Comment**

Review-16917

Official Review by Program Committee QHWY 06 Oct 2025, 16:57 (modified: 08 Nov 2025, 02:05)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee QHWY, Authors
Revisions (/revisions?id=sOASjxlFAJ)

Review:
Strengths:

1. The authors effectively combine historical cases from philosophy, physics, and mathematics to support and contextualize their main argument.
2. The overall exposition is clear.

Weaknesses:

1. The core claim of the paper is that the world model framework is an inadequate theoretical lens for characterizing what understanding entails. Therefore, the authors must clearly define the boundary of what understanding means. The term is too broad and ambiguous in its current form.
2. The discussion mainly focuses on state–transition world models and explicitly excludes video generation models, causal world models, and other modern variants. This limitation should be explicitly stated in the title, abstract, and introduction to avoid potential misinterpretation.
3. Drawing a strong conclusion that "world models ≠ understanding" based on only three philosophical case studies seems overgeneralized. Extending this argument across various domains may not be fully justified.
4. Following the above point, the paper needs empirical or analytical evidence to support its conceptual claims. Currently, it remains entirely theoretical without any experimental validation or case-based analysis.
5. Since world models are generally recognized as frameworks for representing and predicting world states, the paper should further clarify the distinction between understanding and imagination capabilities.
6. If a world model can self-evolve, update its knowledge base, or perform self-correction through feedback, to what extent could such a system be considered to possess a certain level of understanding?
7. As an open discussion point, do large-scale language models, which have massive parameters and emergent reasoning abilities, demonstrate any form of understanding that the authors claim?
8. The paper should provide precise definitions and avoid vague descriptive terms such as "true understanding" or "genuine insight." These phrases make the argument sound rhetorical rather than analytical.

Rating: 5: Marginally below acceptance threshold

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Add: **Ethics Chair Author Comment**

Beyond World Models: Rethinking Understanding in AI Models

Official Review by Program Committee Qnex 06 Oct 2025, 01:42 (modified: 08 Nov 2025, 02:05)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee Qnex, Authors

Revisions (/revisions?id=lmeKbuDhHQ)

Review:

Paper summary:

- The paper asserts that, in the machine learning literature, the world model representation arising in a model is a commonly brought up hallmark of human-level understanding. The paper then disputes that possessing a world model representation constitutes human-like understanding. The paper presents several case studies, illustrating how the presence of the world model representation, construed in a particular way, is insufficient to conclude that the model has achieved what would be considered a human-level understanding of a topic. The paper considers counterarguments to its position – including that the broad view of the of a world model representation with enriched states might resolve the gap between world model representation and understanding – at the cost of the potential falsifiability of the world model framework.

Review summary:

- The paper offers a though-provoking discussion of what is meant by a world model representation and if a world model is sufficient to conclude the presence of a human-level understanding – however, uncrisp definitions of key concepts and some circularity in the core argument (world models are narrowly construed by the authors and for that reason – by design! - are not enough to capture human-like understanding) undermine the message of the work. However, some reframing might alleviate these issues – e.g., refocusing the paper on narrowly vs. broadly construed world models and the meaning of world model representation in the literature. See specific points for the details.

Specific points:

Positives:

- The discussion is overall thought-provoking, attempting to nail down what different authors are trying to get at when advocating for learning that gives rise to the world model representations – and to what degree such representations would be sufficient to conclude the presence of human-like understanding in a model.

Negatives:

- The main and critical issue I have with the paper right now is lack of a crisp definition of what is “human-like understanding” – therefrom arise the problems. One approach in epistemology has been to highlight the distinction between (1) surface-level knowledge of some phenomenon and (2) the true understanding of the phenomenon – which includes the deeper “why” behind why the phenomenon is as it is – which itself is a special kind of knowledge that is not captured in the former knowledge bucket – so, in that framework, understanding is basically some baseline knowledge + additional contextual knowledge. (https://www.tandfonline.com/doi/full/10.1080/0020174X.2022.2146186 (https://www.tandfonline.com/doi/full/10.1080/0020174X.2022.2146186))
- From this standpoint then, the paper seems to argue that the world model representation does not capture all kinds of knowledge usually present in humans. But the reason for why such knowledge is not captured by the world model is precisely because the authors deny the framework the ability to contain such contextual knowledge by construing the framework narrowly, arguing it would otherwise come at a “price of potentially undermining its [world model approach’s] falsifiability.” This itself seems to be a very circular argument – and I don’t find it to be solid / well-developed. For example, why would allowing for a concept of prime to be represented in a world model risk the approach’s falsifiability? So, in view of this, I don’t think “Beyond world models” is a fair framing for the paper – it could, however, lead to a good discussion of narrowly construed vs. broadly construed world models and what is exactly meant by a world model. (“What do we mean by a world model?” could be a more appropriate title.)
- Overall, I understand that there are no universally accepted definitions for world models and understanding - but deciding and settling on some specific definitions - even if controversial to some - seems to be necessary - and the paper would benefit from this greatly - anchoring the discussion.
- As an apparent discrepancy, on p.1, the paper states that “AI models may well develop genuine understanding through mechanisms that go beyond or differ entirely from world models.” – this means world models are neither (1) necessary, nor (2) sufficient for understanding. Yet, all the case studies seem to only address and illustrate the prong #2 - where world models are insufficient for understanding. I don’t think there is an example to support prong #1 – of understanding where world models are unnecessary. Does this mean world models are necessary but not sufficient for understanding? If so, it would be useful to state this explicitly – or to provide a counterexample. This would more clearly and crisply capture the relation of world model representations vs. understanding – as defined in the paper. (Some works seem to argue world models are necessary https://arxiv.org/pdf/2506.01622 (https://arxiv.org/pdf/2506.01622)).
- The paper asserts “It is often argued that since mental world models are an integral component of how humans understand the physical world, the presence of world models in AI models implies human-like understanding capabilities (LeCun 2022; Ng 2023; Mitchell 2025a; Ser et al. 2025). “ – but it is unclear to me if this asserted position in its strong form (mental world models are an integral component of understanding >> detection of world model representation implies full human-like understanding) actually reflects the literature and is supported by the provided citations – it would be useful if the authors could include specific quotes they rely on in these sources.
- The paper briefly mentions causality – but it is unclear what role it plays in relation to world models and understanding – or if it is important at all - it would be useful to clarify this.

Rating: 5: Marginally below acceptance threshold

Confidence: 5: The reviewer is absolutely certain that the evaluation is correct and very familiar with the relevant literature

Add: **Ethics Chair Author Comment**

Review

Official Review by Program Committee vjvP 05 Oct 2025, 22:47 (modified: 08 Nov 2025, 02:05)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee vjvP, Authors

Revisions (/revisions?id=1Z3SJcjBcM)

Review:

Overall Assessment

First and foremost, this is a very interesting and insightful paper. From the perspectives of the philosophy of science and mathematics, the author provides a profound critical analysis of the popular "world model" concept in the current AI field, arguing for its incompleteness on the path toward "human-level understanding." The

paper's core argument, that simulating and predicting physical changes does not equate to truly understanding the world, is highly persuasive and offers valuable points for reflection for researchers in this field.

I am not an expert in philosophy or related fields, so the following comments and questions are offered from the perspective of an AI researcher in the hope of engaging in a deeper discussion with the author.

Discussion and Questions on the Core Argument

I strongly agree with the paper's fundamental premise: merely being able to predict the regular changes of the physical world is not equivalent to achieving human-level understanding. Through well-chosen case studies, the author powerfully reveals the gap between simulation capabilities and deep understanding. However, regarding how world models learn and the nature of their internal representations, I have a few preliminary thoughts I would like to discuss.

1. On the Abstraction of Physical Laws: Internal Representation vs. Human-Interpretable Symbols

Give an example of Newton's discovery of universal gravitation (though my own extension, it is consistent with the spirit of the Bohr's theory example in the paper): before Newton, humans had long observed that objects fall, but only Newton abstracted this phenomenon into the elegant mathematical process of the law of universal gravitation. The author suggests that world models are merely stuck at the stage of simulating a "falling apple" without grasping the underlying mathematical principles.

My question is: Isn't the core task of a world model precisely to learn and internalize the mathematical process that drives the phenomenon?

A world model learns the process of an apple falling from massive amounts of data, and its learned outcome is embodied in an extremely complex neural network. This network is itself a high-dimensional, non-linear mathematical model. Can we consider this model to have functionally "learned" the law of universal gravitation, with the only difference being that it has not expressed this law in a simple, human-interpretable mathematical formula (like

$$F = G\frac{m_1 m_2}{r^2}$$

), but rather in the form of distributed, high-dimensional weight parameters? In other words, must "understanding" necessarily manifest in a human-interpretable symbolic form? Or could a functionally equivalent but formally inscrutable internal model also be considered a form of understanding?

2. On Mathematical Reasoning: From "Verification" to "Creative Problem-Solving"

The paper's case study on "understanding mathematical proofs" is brilliant, clearly distinguishing between "verification" and "understanding." The author posits that world models are limited to the former.

However, recent developments seem to challenge this view. Many top-tier AI models have achieved remarkable results in highly challenging competitions like the International Mathematical Olympiad (IMO). For example, Gemini Deep Think solved five problems using only natural language in 4.5 hours, scoring 35 points. The specific problem-solving process was also made public. These problems stump even the most brilliant human minds each year and often require extraordinary intuition and creative steps to solve. This performance seems to go beyond mere "verification" and enters the realm of "understanding" or even "creation."

As Hu & Shu (2023) discussed in their paper [1], the reasoning process of Large Language Models (LLMs) is effectively realized through world models. Does the performance of these models in math competitions imply that their underlying world models have already developed capabilities beyond mechanical verification?

Conclusion

Once again, I want to affirm that this is an excellent and thought-provoking paper. It accurately points out potential theoretical blind spots in the current pursuit of "understanding" in AI research. The questions I have raised are not intended to refute the paper's conclusions, but also to share my own perspective on world model "understanding".

[1]. Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., & Hu, Z. (2023). Reasoning with language model is planning with world model. arXiv preprint arXiv:2305.14992.

Rating: 5: Marginally below acceptance threshold

Confidence: 2: The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper

Add: **Ethics Chair Author Comment**

Critiquing the World Models Approach to Understanding in AI

Official Review by Program Committee v3ym 22 Sept 2025, 06:04 (modified: 08 Nov 2025, 02:05)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee v3ym, Authors

Revisions (/revisions?id=R1slsSpAs7)

Review:

This paper tackles the question of how to characterise understanding in AI, focusing on world models as a potential theoretical framework. World models are defined as internal representations that simulate aspects of the external world, track entities and states, capture causal relationships, and enable the prediction of consequences, which contrasts with representations based solely on static correlations.

The main aim of the paper is to critique world models as a candidate theoretical approach to understanding in AI. Although the authors agree that world models represent a significant advance compared to mere surface patterns, they argue that such models have limitations and fail to capture human-level understanding across various domains of physical reasoning and problem-solving. Their central claim is that the world-models framework is not an adequate theoretical account of human-like understanding.

To develop this critique, the authors adopt a case-study approach, selecting three examples from different domains: a computer built from falling dominoes, Bohr's atomic theory, and mathematical proofs. They employ thought experiments—most clearly in the domino computer case—to show that in all of these examples, understanding cannot be obtained from the world model by simply tracking its states. This point is especially evident in the case of mathematical proof, where understanding depends on grasping abstract relations.

The paper also considers possible counterarguments, such as extending world models with abstract notions, but concludes that such moves risk undermining falsifiability and leading to circular explanations. Overall, the review highlights the limitations of world models as a comprehensive account of understanding in AI.

Overall evaluation

The topic addressed by the authors both falls within the long philosophical tradition of analysing knowledge and understanding, and at the same time is a topical and hotly debated issue in the AI community. The authors' conclusions are convincing. I also find the choice of case-study examples interesting and enlightening. I would strongly encourage the authors to continue their research in this area, as I regard it as an important topic for the philosophy and ethics of AI, and their ideas as significant for the community.

That being said, the paper suffers from a number of weaknesses. First of all, I would appreciate clearer definitions of the basic concepts: world models, understanding, and especially human-like understanding. I can see why the authors might wish to avoid strict definitions of understanding, but for a better grasp of the scope of their conclusions at least some definition is indispensable.

In their rebuttal response, the authors point out that they deliberately avoided defining "understanding", arguing that any definition would be counter-productive and lead to "false precision" in an unsettled epistemological concept. They cite philosophical arguments (Popper, Ramsey) suggesting that forcing precise definitions on contested concepts can impede progress. It is true that understanding is a complex, debated concept in epistemology. However, avoiding any definition does not eliminate confusion, but rather invites it. If authors are against any definition, at least some sort of *description of what the authors mean by "human-like understanding"* is needed, otherwise readers are left guessing the scope of the term. It would make the critique of world models more concrete: we'd know what criteria specifically are being used to judge whether a world-model-based system demonstrates understanding or not. Without those criteria, the claim "world models fail to capture human-like understanding" risks being unfalsifiable. In other words, if "understanding" remains undefined, one could always claim a given AI lacks true understanding with no clear benchmark for refutation.

The paper doesn't need o provide a perfect or universally accepted definition of understanding, since I doubt that is in any way possible, but it does need to tell us what *intuition about understanding* is being used to evaluate world models. Otherwise, the central claim remains too vague to rigorously assess or build upon.

Furthermore, it is not entirely clear what exactly the authors claim in relation to the examples they provide. For all world-model instantiations of the examples, is it claimed that they fail to characterise understanding? Or is the quantification existential rather than universal? It seems natural that the claim is universal, but it would

strengthen the paper to present this more formally and rigorously. In general, the paper suffers from a lack of clear structure. The points are interconnected, but some key ideas should be highlighted more prominently. The argumentation might also be simplified, since in its current form it can at times be confusing. A particularly important point—the example of mathematical proofs showing that formal derivation allows for verification but does not provide understanding—should be emphasised and explained further, as it is crucial for the overall argument.

In addition, it seems that the abilities which world models lack, such as the capacity to identify key or novel points, and especially the ability to motivate the proof—that is, to explain why certain steps are natural or to be expected—are not immediately clear in terms of how they could be measured. Thus, while the authors criticise the world-model framework, it is left somewhat unclear what alternative framework might be proposed, and whether such an alternative could be empirically tested.

Finally, the question of why human-like understanding is desirable to measure is not addressed anywhere. It would be valuable to provide some motivation for why this matters for AI, since the philosophical question of the nature of understanding can, in principle, be studied independently of AI.

Overall, I think the paper raises important and interesting questions, and the approach the authors take is both promising and reasonable. However, I would strongly encourage them to improve the readability and clarity of their arguments and claims. I would be very interested in seeing this line of research developed further.
Rating: 6: Marginally above acceptance threshold
Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Add: **Ethics Chair Author Comment**

NOT Sure about the quality of philosophical paper

Official Review by Program Committee L8Kx 01 Sept 2025, 19:04 (modified: 08 Nov 2025, 02:05)
Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee L8Kx, Authors
Revisions (/revisions?id=wIKc24iCXV)

Review:
This paper focuses on the concept of world models, and uses problems and case studies from philosophy of science literature to critically examine whether the world framework adequately characterizes human-level understanding. It focuses on specific philosophical analyses where the distinction between world model capabilities and human understanding is most pronounced. The manuscript examines three cases from philosophical work, including 1) Hofstadter's analysis of a computer built from falling dominoes; 2) Popper's account of understanding physical theories through their problem situations; and 3) Poincare's distinction between verifying and understanding mathematical proofs.

Detailed concerns:

1. I am a researcher in embodied AI, therefore, it is hard for me to understand the philosophical concept in this paper. To my knowledge, an AAAI paper should be a technical paper, that contains a clear motivation, a novel method, and comprehensive experimental results. I am not sure such a philosophical paper (only 5 pages of analysis, without experiments) meets the acceptance requirements of an AAAI paper.
2. As for the definition of world model, the concept of world model in this paper is a little different from the general concept in the field of embodied AI and cognitive neuroscience. In this paper, world model means just internal representations that simulate aspects of the external world, track entities and states, capture causal relationships, and enable the prediction of consequences. While in embodied AI, the world model focuses on modeling the environment state and the transition between states. Besides, the world model in embodied AI is not only for understanding, but also for improving the training efficiency and effectiveness of the embodied policy. These two concepts are a little different.
3. In the related work, many existing world model works, including diffusion-based, video-generation-based, and 3DGS-based methods, are not discussed.

Rating: 5: Marginally below acceptance threshold
Confidence: 2: The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper

Add: **Ethics Chair Author Comment**

Interesting and insightful work, but no empirical evidence.

Official Review by Program Committee 13rf 29 Aug 2025, 23:15 (modified: 08 Nov 2025, 02:05)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee 13rf, Authors

Revisions (/revisions?id=3YHGHBJYFo)

Review:

Summary

The paper argues that current world models cannot capture human-level understanding, drawing on three philosophical case studies: (1) a domino computer (missing abstract concepts like primality), (2) Bohr's atomic theory (missing problem-context and motivation), and (3) mathematical proofs (missing the strategy and insight behind proof steps). The perspective is interesting but lacks empirical validation.

Strength

1. Interesting perspective of understanding the current world models.
2. The paper is well-written and engaging, I really enjoyed reading it.
3. The case studies effectively highlight the conceptual gaps between what we need and what current world models can do.

Weakness

1. The paper is entirely philosophical, with no empirical experiments.
2. It would strengthen the contribution if computational experiments can be added.

Rating: 5: Marginally below acceptance threshold
Confidence: 3: The reviewer is fairly confident that the evaluation is correct

Add: **Ethics Chair Author Comment**

AI Review

AI Review by Program Committee AI 27 Aug 2025, 12:26 (modified: 11 Oct 2025, 00:29)

Program Chairs, Area Chairs, Senior Program Committee, Program Committee Submitted, Program Committee AI, Authors

Revisions (/revisions?id=houis0QGwF)

Review:

Title: Beyond World Models: Rethinking Understanding in AI Models

Synopsis of the paper The paper critiques the inference from "internal world-model-like representations" in AI systems to human-like understanding. Through three case studies—a domino-based primality computer, Popper's notion of understanding a theory via its problem situation, and Poincaré's distinction between proof verification and understanding—it argues that state-and-transition modeling misses explanatory, contextual, and strategic aspects central to understanding. It anticipates the counterargument that world models can be extended to include abstract states and contends that such moves risk unfalsifiability. The contribution is conceptual rather than empirical, aiming to sharpen how "understanding" is evaluated in AI.

Summary of Review This is a clear, well-argued conceptual critique that addresses a prominent line of reasoning in contemporary AI: that internal state-and-transition "world models" suffice for understanding. The case studies are apt and persuasively separate predictive simulation from explanatory abstraction, problem-situation grasp, and proof-level insight. The main limitations are definitional narrowness (limited engagement with hierarchical/causal formulations under which "world models" make falsifiable commitments), lack of operationalization into testable evaluations, and several correctable presentation issues. With sharpened scope, engagement with work that constrains internal structure causally, and concrete evaluation protocols targeting the highlighted aspects of understanding, the paper could be an influential conceptual reference.

Strengths

- Clear targeting of a live claim: The paper pinpoints the often-implicit inference from internal state-and-transition representations to understanding and interrogates its validity rather than disputing that AI can ever understand.
- Apt philosophical case studies with concrete traction for AI:
  - Domino computer: Cleanly illustrates that mechanistic state-tracking explains proximate causes (dominoes' falls) but not the abstract explanans (primality) driving the system's behavior.
  - Problem-situations (Popper): Emphasizes that understanding a theory requires situating it within a landscape of problems and explanatory aims, beyond picturing its internal dynamics.
  - Proofs (Poincaré, Avigad): Distinguishes verification from understanding of strategy, key ideas, and motivation, aligning with practical challenges in aligning model rationales with genuine reasoning.
- Anticipation of the "just add abstraction" response: The paper correctly notes that enriching state spaces without constraints invites unfalsifiable, post hoc reinterpretation, reducing explanatory value.
- Relevance to ongoing empirical narratives: The critique productively contextualizes celebrated findings of internal state tracking in structured domains (e.g., games) and world-model claims for video/agent systems, discouraging over-interpretation as understanding.

Weaknesses

- Definitional narrowness and limited engagement with stronger formulations:
  - The critique focuses on world models as state-and-transition tracking but only briefly engages with formulations that explicitly align internal structure with high-level variables and counterfactual behavior, which yield falsifiable commitments. Interchange intervention training induces mappings between internals and causal variables (Geiger et al., 2022), and recent counterfactual formulations for language models move toward structural equation modeling (Ravfogel et al., 2025). The paper would benefit from analyzing whether such approaches mitigate the unfalsifiability it highlights, and why they still fall short of the paper's understanding criteria.
  - Mechanistic interpretability advances (e.g., gated sparse autoencoders that find sparse, semantically coherent features) offer a concrete path to test whether abstract properties can be internalized without arbitrary retrofitting (Rajamanoharan et al., 2024). This line is not discussed.
- Lack of operationalization and evaluation protocols:
  - No concrete tests or metrics are proposed for the identified aspects of understanding: recognizing abstract explanantia across different physical realizations, articulating problem-situations and explanatory aims, or identifying key insights and high-level proof strategies. Without an evaluation blueprint, it is hard to convert the critique into cumulative empirical progress.
  - Existing empirical work provides actionable precedents: unfaithful chain-of-thought rationales (Turpin et al., 2023), symbolically grounded and verifiable reasoning (Xu et al., 2024), and structure-imposing proof generation (Zheng et al., 2024). These could be leveraged to turn the philosophical distinctions into measurable desiderata.
- Insufficient separation of prediction/control competence from understanding:
  - Evidence that sequence models linearly encode environment state and enable causal interventions (Li et al., 2023) or perform well in richer domains (Feng et al., 2023), and that world models improve long-horizon memory/control (Samsami et al., 2024), does not by itself establish understanding as described in the paper. The critique asserts this distinction but could more explicitly frame evaluation criteria that go beyond performance and simulation fidelity.
- Technical and presentation issues (all correctable, do not undermine the thesis):
  - In Section 3.3, the Euclid example states "Consider $N + 1$, where N is the product of all primes on our list," while Appendix C defines N as the product-plus-1; this is inconsistent and may confuse readers.
  - Figure 2 labels the split "N is either prime or composite" as "excluded middle." A neutral "case split" label would avoid suggesting a stronger logical principle than needed.
  - The sentence linking "Turing's theory" to Curry–Howard overstates the relationship; Curry–Howard is a correspondence between proofs and typed programs, not an identity claim "in Turing's theory."
  - Appendix D: the piecewise definition of f must present triples in each branch (e.g., $(x + 2z, z, y - x - z)$ in the first branch); any omission would make f ill-defined on S and invalidate the involution/parity reasoning. It would help to add a brief justification that f is well-defined and involutive on S before invoking the parity principle.
  - The reproducibility checklist marks the presence of formal theorems/proofs and experiments; the main text introduces no novel formal theorems, and the paper is conceptual. The checklist should be updated to reflect the paper's nature.

Suggestions for Improvement

- Clarify scope and constraints:
  - Add a short “Scope and Definitions” subsection that precisely defines the target notion of world models (e.g., internal representations that track entities, discrete states, and causal transitions learned from data) and what is out of scope (e.g., arbitrary post hoc abstraction layers not tied to falsifiable commitments).
  - State explicit criteria under which adding abstract states is acceptable without collapsing into unfalsifiability (e.g., abstractions must support interventional predictions, cross-substrate transfer, or satisfy structural constraints that distinguish competing abstractions).
- Engage with falsifiable causal/mechanistic approaches:
  - Discuss interchange intervention training and related techniques that align internal representations with an external causal model (Geiger et al., 2022). Analyze whether such methods address the critique’s core concern or why they remain insufficient for the paper’s understanding criteria (e.g., they can align variables but still do not explain problem-situations or strategy).
  - Connect to mechanistic feature discovery (Rajamanoharan et al., 2024) to delineate when discovered features count as abstract explanantia versus post hoc naming.
  - Situate the critique alongside state-encoding successes (Li et al., 2023; Feng et al., 2023) and world-model performance gains in control/memory (Samsami et al., 2024), explicitly separating predictive/control competence from the paper’s targeted dimensions of understanding.
- Propose concrete evaluation protocols aligned with the paper’s desiderata:
  - Explanatory abstraction across realizations: Construct multi-realization benchmarks where the same abstract property governs different physical implementations (dominoes, circuits, cellular automata). Evaluate whether models can (i) identify the abstract property (e.g., primality), (ii) explain system behavior in terms of it, and (iii) transfer explanations across substrates.
  - Problem-situation understanding: Design tasks requiring articulation of the problem a theory was introduced to solve (e.g., why Bohr posits quantized orbits to explain discrete spectra), with rubrics grounded in philosophy of science. Include contrastive questions (“why this assumption rather than that?”) to elicit explanatory commitments.
  - Proof understanding beyond verification: Build benchmarks that require identifying key insights, motivating steps, and providing a high-level strategy outline, scored against criteria inspired by Avigad’s framework. Leverage faithful/symbolic chain-of-thought (Xu et al., 2024) and structure-aware proof generation (Zheng et al., 2024), and use unfaithfulness tests (Turpin et al., 2023) to guard against spurious rationales.
  - Counterfactual competence: Incorporate text-level structural equation modeling and counterfactual generation (Ravfogel et al., 2025) to probe causal reasoning beyond picturing/simulation, aligning with the paper’s emphasis that “picturing is not understanding.”
- Strengthen argumentation around the counterargument:
  - Refine Section 4 by articulating non-circular acceptance criteria for enriched world models (e.g., interventional adequacy, cross-domain generality, explanatory parsimony). This will turn the unfalsifiability concern into constructive guidance for when abstraction enhances, rather than evacuates, explanatory value.
- Correct and tighten presentation:
  - Harmonize the Euclid notation (N vs. $N+1$) across the paper.
  - Relabel “excluded middle” to “case split” in Figure 2.
  - Rephrase the Curry–Howard sentence to: “There is a close correspondence between proofs and programs, famously captured by the Curry–Howard correspondence.”
  - In Appendix D, ensure the piecewise definition of f is fully specified with triples in each branch, and add a brief verification that f is an involution on S before applying the parity argument.
  - Update the reproducibility checklist to reflect that the paper is conceptual, with no new formal theorems or experiments.

References

- Feng, X., Luo, Y., Wang, Z., Tang, H., Yang, M., Shao, K., Mguni, D., Du, Y., & Wang, J. (2023). ChessGPT: Bridging policy learning and language modeling. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track. https://proceedings.neurips.cc/paper_files/paper/2023/hash/16b14e3f288f076e0ca73bdad6405f77-Abstract-Datasets_and_Benchmarks.html

(https://proceedings.neurips.cc/paper_files/paper/2023/hash/16b14e3f288f076e0ca73bdad6405f77-Abstract-Datasets_and_Benchmarks.html)
- Geiger, A., Wu, Z., Lu, H., Rozner, J., Kreiss, E., Icard, T., Goodman, N., & Potts, C. (2022). Inducing causal structure for interpretable neural networks. Proceedings of the 39th International Conference on Machine Learning (ICML 2022), Proceedings of Machine Learning Research, 162, 7324–7338. https://proceedings.mlr.press/v162/geiger22a.html (https://proceedings.mlr.press/v162/geiger22a.html)
- Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2023). Emergent world representations: Exploring a sequence model trained on a synthetic task. In International Conference on Learning Representations (ICLR 2023). https://iclr.cc/virtual/2023/oral/12641 (https://iclr.cc/virtual/2023/oral/12641)
- Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., & Nanda, N. (2024). Improving sparse decomposition of language model activations with gated sparse autoencoders. In Advances in Neural Information Processing Systems 37 (NeurIPS 2024). https://papers.nips.cc/paper_files/paper/2024/hash/01772a8b0420baec00c4d59fe2fbace6-Abstract-Conference.html (https://papers.nips.cc/paper_files/paper/2024/hash/01772a8b0420baec00c4d59fe2fbace6-Abstract-Conference.html)
- Ravfogel, S., Svete, A., Snæbjarnarson, V., & Cotterell, R. (2025). Gumbel counterfactual generation from language models. In International Conference on Learning Representations (ICLR 2025). https://proceedings.iclr.cc/paper_files/paper/2025/hash/e15790966a4a9d85d688635c88ee6d8a-Abstract-Conference.html (https://proceedings.iclr.cc/paper_files/paper/2025/hash/e15790966a4a9d85d688635c88ee6d8a-Abstract-Conference.html)
- Samsami, M. R., Zholus, A., Rajendran, J., & Chandar, S. (2024). Mastering memory tasks with world models. In International Conference on Learning Representations (ICLR 2024). https://proceedings.iclr.cc/paper_files/paper/2024/hash/0a0e436e8830f746300f592377730fca-Abstract-Conference.html (https://proceedings.iclr.cc/paper_files/paper/2024/hash/0a0e436e8830f746300f592377730fca-Abstract-Conference.html)
- 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. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023). https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html (https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html)
- Xu, J., Fei, H., Pan, L., Liu, Q., Lee, M.-L., & Hsu, W. (2024). Faithful logical reasoning via symbolic chain-of-thought. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), 13326–13365. https://doi.org/10.18653/v1/2024.acl-long.720 (https://doi.org/10.18653/v1/2024.acl-long.720)
- Yang, K., Swope, A., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R. J., & Anandkumar, A. (2023). LeanDojo: Theorem proving with retrieval-augmented language models. In Advances in Neural Information Processing Systems 36 (NeurIPS 2023). https://nips.cc/virtual/2023/oral/73738 (https://nips.cc/virtual/2023/oral/73738)
- Zheng, Z., Malon, C., Min, M. R., & Zhu, X. (2024). Exploring the role of reasoning structures for constructing proofs in multi-step natural language reasoning with large language models. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), 15299–15312. https://doi.org/10.18653/v1/2024.emnlp-main.854 (https://doi.org/10.18653/v1/2024.emnlp-main.854)

(/))