Beyond World Models: Rethinking Understanding in AI Models

Tarun Gupta, Danish Pruthi

Indian Institute of Science
Bengaluru, KA, India
{tarungupta, danishp}@iisc.ac.in

Abstract

World models have garnered substantial interest in the AI community. These are 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 that humans possess such mental world models, and finding evidence of similar representations in AI models might indicate that these models "understand" the world in a human-like way. In this paper, we use 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 help us explore the limits of world models.

1 Introduction

In artificial intelligence, the concept of world models raises fundamental questions across domains: Do LLM representations track world states and the transitions between them, and do they use these representations to predict next tokens? Do video-generation models create representations of physical laws and spatial geometry, predicting future frames by simulating these learned laws of nature? At its core, the world model hypothesis asks whether neural networks capture and reproduce the actual causal processes that generated their data, or whether they merely manipulate surface patterns and capture correlations without intermediate representations that mirror real-world mechanisms (Andreas 2024).

The motivation for studying world models stems from the human experience of mental visualization and picturing, along with our ability to mentally simulate these visualized mental models. A quintessential example is the heliocentric model of the solar system, where humans visualize the sun, planets, and other celestial bodies as entities with specific states (positions, velocities) that transition according to physical laws governing their orbits.[1]

This intuitive appeal of world models raises the question: If AI models (e.g., LLMs) can maintain such world states and model state transitions rather than just leveraging surface-level correlations, would this constitute human-like understanding? 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 suggests human-like understanding capabilities (LeCun 2022; Ng 2023; Mitchell 2025a; Ser et al. 2025).[2] While both world models and understanding lack universally agreed-upon definitions, the growing interest in world models within AI research makes this question important to examine. In this paper, we argue that while world models represent a crucial advance beyond mere surface patterns, they fail to capture human-level understanding across various domains of physical reasoning and problem-solving.

To be clear, our argument is not that AI models cannot achieve understanding. We do not claim that current or future AI models lack the capacity to understand the world. Instead, we critique the world model framework as an inadequate theoretical lens for characterizing understanding. AI models may well develop understanding through mechanisms that go beyond or differ entirely from world models. Our critique targets the specific claim that possessing world model-like representations constitutes understanding.

Rather than attempting to provide universal criteria for understanding across all domains, we adopt a case study approach that examines particular, yet fundamental, instances of understanding where the limitations of the world model framework become apparent. Understanding is multifaceted—in any given domain, there are multiple valid perspectives on what constitutes understanding. Our strategy is deliberately selective: we focus on particular philosophical analyses that illustrate aspects of understanding that go beyond what world models can capture. We acknowledge that other perspectives on understanding might align better with world models. However, our selected cases help us explore limitations of using world models to characterize human-level understanding.

To this end, we examine three cases from philosophical

[1]While one commonly speaks of mental models of the 'real world,' these are more precisely models of our theories about the world, such as the heliocentric model itself, or even superseded theories like the geocentric model with its epicycles, or Bohr's model of electrons orbiting the atomic nucleus in discrete paths.

[2]See §E for specific quotations from these cited works on the connection between world models and understanding.

work: (1) Hofstadter's (2007) analysis of a computer built from falling dominoes, (2) Poincaré's (1914) distinction between verifying and understanding mathematical proofs, and (3) Popper's (1979) account of understanding physical theories through their problem situations. Through these case studies, we show that the world-model conception fails to capture what these philosophical analyses reveal about human-level understanding. Our analysis contributes to discussions about world model research and its theoretical foundations.

2 Background and Related Work

World Models in AI. The term "world model" draws considerable attention in the AI community and is considered a key ingredient for building general intelligence (LeCun 2022; Ding et al. 2024). However, it lacks a universally-accepted definition. Different researchers define world models differently (Ding et al. 2024; Xing et al. 2025). For this paper, we define world models—following the prevailing conception—as internal representations that track objects, their states, and the rules governing how states change over time.

To investigate whether AI models develop such world models, researchers use probing techniques to examine internal representations in neural networks. This involves analyzing learned features in specific layers, studying activation patterns, or using linear decoding methods to recover world state representations from the model's internal states. These probing approaches aim to determine whether models maintain interpretable representations of world states and transitions rather than relying solely on surface-level pattern matching. For LLMs, Li et al. (2023) achieve a landmark result using such probing techniques, showing that a language model trained to play the board game Othello developed an internal world model of the game. Specifically, they find that the model's internal representations can be linearly decoded to recover the actual board state at each move, suggesting the model maintains an internal simulation of game states and transitions rather than relying solely on surface-level pattern matching in move sequences. Similar evidence of internal world models has been found in LLMs trained on chess (Karvonen 2024). On the other hand, some research suggests these world models are not clean, human-like mental models, but collections of learned heuristics (Karvonen et al. 2024; Nikankin et al. 2024). For an excellent discussion of this topic, we refer readers to (Mitchell 2025b).

Advances in multimodal models have expanded the use of the term "world models" to include video generation systems. Models like Sora (OpenAI 2024), WorldGPT (Yang et al. 2024) and Veo (Google DeepMind 2025) are called world models because they generate videos that appear to follow physical laws and maintain temporal consistency. However, this capability to produce visually plausible sequences is distinct from the question of whether a model's internal representations track discrete states and model transitions between them. Likewise, other categories of models, including gaming world models (Bruce et al. 2024), 3D scene models (World Labs 2025) and physical-world models (Agarwal et al. 2025), are also sometimes termed world models, each with different definitions and capabilities (Xing et al. 2025). Such models are not relevant to our discussion, which examines whether the presence of world model-like representations that explicitly track states and transitions in AI models constitutes human-like understanding.

Philosophical Perspectives on Understanding. "Understanding" has been a subject of intense debate in philosophy of science and philosophy of mathematics. For a general overview of the literature on understanding, see (Grimm 2011, 2017; Baumberger, Beisbart, and Brun 2016). For discussions specifically focused on mathematical understanding, see (Avigad 2008; Hamami and Morris 2024). Explanation and understanding are closely related, where the latter is seen as the goal of the former (Friedman 1974; Grimm 2010).

Given the numerous theories and frameworks proposed to explain the nature of understanding, arriving at a single, unified account is challenging. Therefore, we avoid defining "understanding" in this paper, as providing a definition—even an operational one—would be counter-productive. Understanding remains a contested concept in epistemology without agreed definition, and imposing one would produce only false precision. Any operational definition would merely shift the problem to the defining terms, leading to an infinite regress unless we admit primitive terms, that is, undefined terms (Popper 1945). Rather than impose an arbitrary definition on a concept 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 conception of understanding falls short of explaining human-level understanding.

3 Case Studies: Understanding Beyond World Models

This section examines three cases where the world model conception falls short of characterizing human-level understanding: a computer built from falling dominoes (§3.1), mathematical proofs (§3.2), and Bohr's atomic theory (§3.3).

3.1 Understanding a Computer Built from Dominoes

The prevailing conception in world model research holds that understanding physical scenes involves maintaining states for discrete, recognizable objects and tracking the intuitive physics relationships between them. These states, which researchers often probe for in the internal representations of models, correspond to what seems meaningful to humans as states. For example, in board games, each square represents a distinct state rather than the microscopic wood grain patterns or exact molecular arrangements of the board material. Similarly, in video understanding, states might track objects like cars or people rather than individual pixel intensities or spectral frequencies. However, when we apply this approach of state selection to Hofstadter's (2007) thought experiment, we see it can miss understanding a system's behavior.

In this thought experiment, dominoes are arranged into a mechanical computer that determines whether numbers are prime. When a domino falls, it pops back up after a fixed time, thereby propagating signals along carefully-arranged networks. With such a system, we can implement a mechanical computer where signals travel down stretches of dominoes

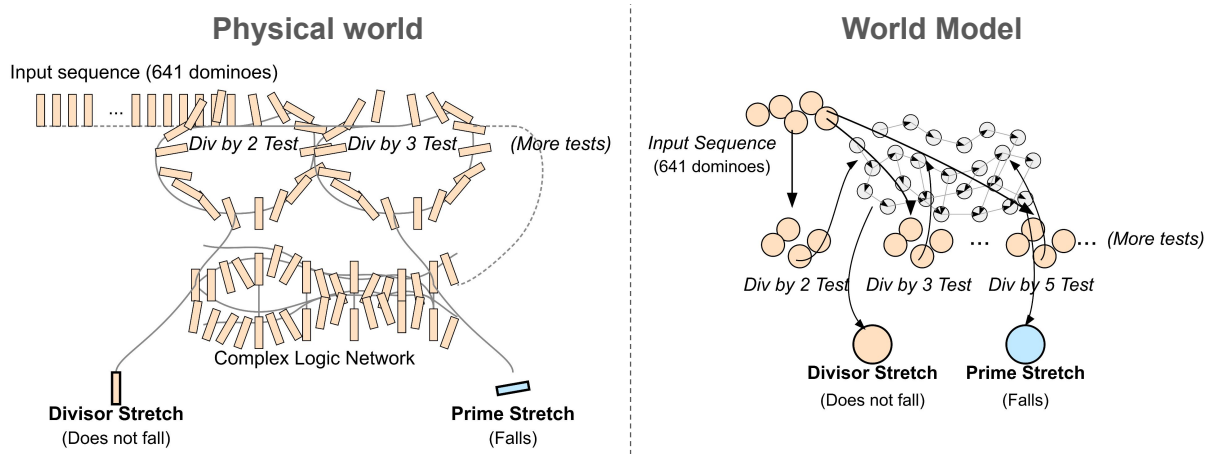

Figure 1: Left: Conceptual illustration of physical arrangement of dominoes in a computational system (Hofstadter 2007). Right: A schematic world-model representation showing states and causal relationships between dominoes. While the world model can track physical states (standing or fallen dominoes) and predict how one domino causes another to fall, it fails to capture the abstract mathematical concept of primality that fundamentally explains the system's behavior.

that bifurcate, join together, propagate in loops, and jointly trigger other signals. Relative timing is of course crucial, but the specific implementation details are not relevant to our discussion. The basic idea is that a precisely arranged network of domino chains can function as a computer program for carrying out computations—in this case, determining if a number is prime. (See §A for a detailed description of this mechanical computer.)

To test primality, the system takes input by placing exactly that many dominoes (e.g., 641) end-to-end in a designated "input stretch." When triggered, the system runs various tests for divisibility by potential factors. If any divisor is found, a signal travels down a specific "divisor stretch," indicating the number is not prime. Conversely, if no divisors are found, a signal travels down a "prime stretch," confirming primality. This arrangement is shown schematically in Figure 1.

A world model approach to understanding this system would track each domino's position (standing or fallen) at each moment and simulate the physical propagation of falling patterns. When examining why a particular domino never falls when the input is 641, such an approach would focus on the immediate physical causes: none of its neighboring dominoes ever fall. This answer, while physically accurate, merely shifts attention to other dominoes. Tracing backward through the causal chain would eventually lead to a statement of the kind: "That domino did not fall because none of the patterns of motion initiated by the first domino ever include it." This mechanistic tracking of states fails to capture understanding of this system's behavior. The key to understanding lies in recognizing that 641 is prime, an abstract mathematical property that explains the entire pattern of domino behaviors. This understanding cannot be obtained by simply tracking the states of dominoes falling or not falling—no amount of state tracking can reveal the fundamental mathematical concept of primality that governs the system's behavior.

A natural response to this analysis might be that world models could be enriched to include abstract concepts like primality as states themselves—perhaps representing "641 is prime" as a state connected to the physical domino configurations. While such an approach would address this critique, it would come at the price of reducing the framework's explanatory power and limiting its falsifiability. We address this counterargument and its broader implications in §4.

3.2 Understanding Mathematical Proofs

In Turing's theory, computations are essentially the same thing as proofs: every valid proof can be converted to a computation that computes the conclusion from the premises, and every correctly executed computation is a proof that the output is the outcome of the given operations on the input. This fundamental equivalence, formalized in the Curry-Howard correspondence (Howard et al. 1980), establishes that mathematical proofs are sequences of logical transformations that can be viewed as physical processes: step-by-step manipulations of symbolic expressions according to formal rules. World models, though usually conceived for understanding aspects of physical reality, are equally applicable to proof-understanding given that proofs constitute physical processes of symbolic manipulation. Hu and Shu (2023) also argue that world models are important for mathematical reasoning more broadly, suggesting that explicit modeling of intermediate mathematical conclusions and internal simulation of future states is required for mathematical reasoning. Given this connection between world models and mathematical reasoning, and proof-understanding being a subset of mathematical reasoning with extensive philosophical literature (Avigad 2008; Hamami and Morris 2024), proof-understanding offers a focused domain for examining whether world model approaches adequately capture human-level understanding.

A world model approach to proof-understanding would treat proofs as sequences of state transitions, tracking the logical validity of each step. When asked why a particular conclusion holds, such a model would trace backward through the chain of logical states. For example, consider

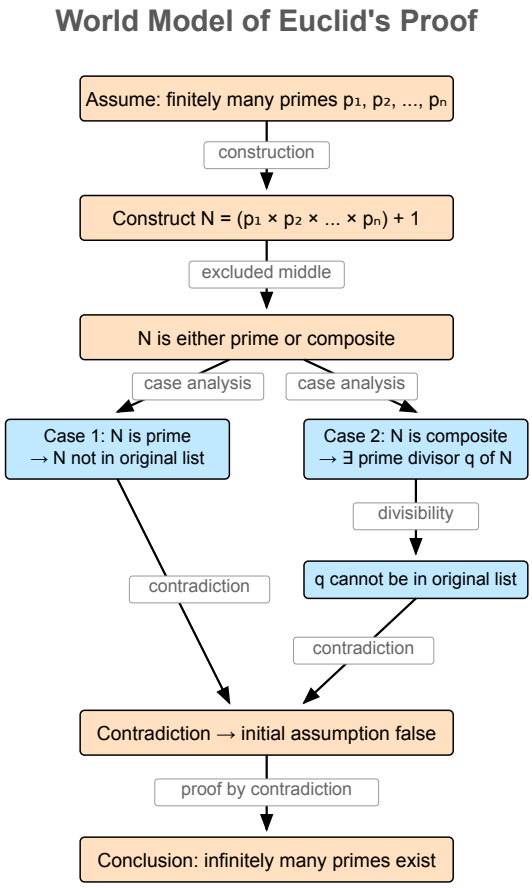

Figure 2: A world model representation of Euclid's proof showing logical states and transitions between them.

Euclid's famous proof that there are infinitely many primes (Heath et al. 1956) (see §C for the proof). Figure 2 illustrates how a world model would represent this proof as a sequence of logical state transitions. The final state might show "Therefore, there are infinitely many prime numbers." The immediately preceding state might contain "Since our assumption led to a contradiction, the original assumption that there are finitely many primes must be false." Working backwards, we might find "Consider $N + 1$, where N is the product of all primes on our list." This approach amounts to mere verification—confirming that each state transition (logical step) adheres to the rules of logical deduction and that the chain of states connects the premises to the conclusion. But does such verification constitute human-like understanding? No. It's a common observation in mathematics that there is an important difference between understanding a proof and verifying it. As Poincaré (1914) observes[3]:

> *Does understanding the demonstration of a theorem consist in examining each of the syllogisms of which it is composed in succession, and being convinced that it is correct and conforms to the rules of the game? [...]*
>
> *Yes, for some it is; when they have arrived at the conviction, they will say, I understand. But not for the majority. Almost all are more exacting; they want to know not only whether all the syllogisms of a demonstration are correct, but why they are linked together in one order rather than in another. As long as they appear to them engendered by caprice, and not by an intelligence constantly conscious of the end to be attained, they do not think they have understood*

For an agent with a world-model conception of understanding, the state transitions in a proof appear as if "engendered by caprice." While world models may help in simulating future states of proof steps—as suggested by (Hu and Shu 2023) to be important for reasoning—this capability alone cannot explain "why [the syllogisms of the proof] are linked together in one order rather than in another."

To illustrate this gap more dramatically than the Euclid's proof example, consider Zagier's (1990) one-sentence proof that every prime $p \equiv 1 \pmod 4$ is a sum of two squares (see §D for brief exposition of the proof). Verifying this proof requires only basic understanding of set properties, involution, and fixed-points. Each step can be checked for logical validity in a straightforward manner. However, such verification is not sufficient for understanding this proof (see (Karai 2010) on how many mathematicians can easily verify Zagier's (1990) proof but struggle to understand it).

To see why, we consider some abilities required to demonstrate proof-understanding from Avigad's (2008) framework (other criteria, for example those proposed by Hamami and Morris (2024), would similarly demonstrate the limitations).

Consider the ability to "indicate 'key' or novel points in the argument, and separate them from the steps that are 'straightforward' ": a world model can verify that the involution f has exactly one fixed point, but it cannot identify why this step is the crucial insight versus the more routine verification that f is indeed an involution. Similarly, for the ability to "'motivate' the proof, that is, to explain why certain steps are natural, or to be expected": a world model cannot explain why constructing the specific set $S = \{(x, y, z) \in \mathbb{N}^3 : x^2+4yz = p\}$ was a natural choice, or why applying an involution to count fixed points would lead to the desired conclusion. Lastly, let us consider the ability to "give a high-level outline, or overview of the proof": while a world model can track the sequence of logical steps, it cannot provide the overarching strategy—that the proof works by cleverly counting the same set in two different ways using properties of involutions. Likewise, many other abilities posed in (Avigad 2008) present similar challenges to the world model conception of understanding.

3.3 Understanding Physical Theories

Understanding a physical theory requires more than simulating its internal mechanisms or predicting its outcomes—it involves understanding the problem situation that led to proposing that theory as a solution. This perspective, developed by Popper (1979), defines "problem situation" as not

[3]This distinction is widely remarked upon, not just by Poincaré (1914); for collections and philosophical discussion of many such statements by mathematicians, see (Avigad 2008; Hamami and Morris 2024).

only the problems one tries to solve but also the theoretical landscape that necessitated the solution—the inadequacies of existing theories and the specific explanatory gaps that generated new problems requiring solutions. In this perspective, understanding includes grasping the explanatory structure that motivated the theory's construction—why certain theoretical commitments were necessary to solve the identified problems and what makes this particular solution explanatorily adequate.

Consider Bohr's atomic theory (Bohr 1913)[4]: Bohr proposed that electrons orbit the nucleus in discrete, fixed energy levels rather than in continuous paths as in classical physics. The key to understanding it is not merely visualizing, or world-modeling electrons jumping between orbits, but recognizing what Bohr attempted to explain with these electron jumps: the sharp, discrete spectral lines observable in atomic spectra. To explain these definite, discrete lines, Bohr had to assume certain discreteness in electron movement possibilities, leading to the concept of jumps between tracks.

Crucial to this explanation is the energy transfer mechanism Bohr proposed: when an electron jumps from an outer orbit to an inner orbit, the atom loses energy, which is emitted in the form of light radiation. The specific frequencies of light observed in spectral lines correspond directly to the energy differences between the allowed electron orbits. This mechanism explains why spectral lines appear at precise frequencies rather than a continuous spectrum.

Without knowing why Bohr introduced this somewhat unnatural model—to explain discrete spectral lines—one cannot understand his theory as a solution to a specific problem situation. The apparent unnaturalness of electrons being constrained to certain orbits and making quantum jumps between them only makes sense in light of the problem Bohr was solving. As Popper (1979) notes, someone who is just presented with the Bohr theory, without knowing that the theory was invented in order to explain the phenomenon of discrete spectral lines, will simply not understand the theory as a solution of a certain problem situation.

The world model alone (electrons orbiting in discrete paths) doesn't capture the understanding of the theory. World models in AI similarly emphasize internal simulation—like picturing electrons on orbits—but as this case-study shows, picturing is not understanding. An AI model might successfully simulate atomic transitions without grasping their importance in broader theoretical context, just as a human might visualize Bohr's orbital structure without understanding its explanatory role in solving the spectral line problem.

4 A Possible Counterargument

A plausible counterargument can be proposed that the issue lies not with world models themselves, but with the level of abstraction we chose for states in our case studies. In our three case-studies, we selected states that align with prevailing conceptions in the literature—tracking objects like dominoes or electrons in the physical world. For mathematical reasoning, we likewise followed approaches like those of (Hu and Shu 2023), which represent the prevailing conception in applying world models to this domain. The counterargument suggests that world models could instead incorporate psychological or social abstractions as states themselves. For instance, in the domino computer example, the concept of primality and the statement "641 is prime" could be represented as states connected to the physical domino configurations. Similarly, for Bohr's theory, discrete spectral lines could be represented as states mapped to the mental picture of electrons orbiting in discrete paths.

While incorporating ad-hoc, abstract states connected with equally abstract transitions would rescue the world model approach from our critique, it would do so at the price of potentially undermining its falsifiability. If states can encode arbitrarily rich abstractions (mathematical properties, problem-solving strategies, historical context, explanatory motivations) then any phenomenon can be retrofitted into the world model approach simply by defining appropriate abstract states and transitions between them. The approach risks becoming irrefutable: any failure to capture understanding can be addressed by adding more complex states, and virtually any cognitive phenomenon can be accommodated by sufficiently enriching the state representations.

This flexibility limits the explanatory contribution of the world model concept itself. If the explanatory work is done by the state representations rather than the world model's dynamics, then the world model approach provides limited theoretical insight beyond organizing existing knowledge into states at various levels of abstraction and connecting them with equally abstract transitions. The counterargument essentially reduces to: "world models can capture understanding if we put understanding into the states"—a circular explanation that presupposes what it seeks to explain.

5 Conclusion

In this paper, we argued that world models fall short of adequately characterizing human-level understanding. We supported this claim through case studies of particular, yet fundamental instances of understanding. Although world models represent an advance over surface-level correlations, our case studies demonstrate important limitations in using world models as a lens for understanding. The world model research program represents, among current approaches, the most significant component, and the precursor to future theories of machine understanding. Our analysis offers only one perspective on how this research direction might be refined. The growing interest in world models as a path to artificial general intelligence makes it valuable to scrutinize whether this framework adequately captures what understanding entails. Philosophical examination of foundational concepts like understanding can complement algorithmic and experimental advances in addressing conceptual limitations of theoretical frameworks. While our perspective may be different from current trends in world model research, we believe such analysis contributes to ongoing discussions about the nature of understanding in artificial intelligence.

[4]For a refresher of Bohr's theory, see §B.

A Hofstadter's Domino Chainium

Hofstadter (2007) introduces the thought experiment of a "domino chainium"—a computer built from dominoes with special properties. In this system, each domino is spring-loaded so that after being knocked down, it automatically returns to its upright position after a short "refractory" period. This enables signals to propagate through the system repeatedly, supporting complex computational processes. The domino chainium functions as a mechanical computer where signals travel through carefully arranged networks of dominoes, bifurcating, joining, and propagating in loops to implement computer programs. The precise timing of domino falls determines how signals propagate and interact throughout the network.

In Hofstadter's example, this system is specifically designed to determine whether a number is prime. To test if a number is prime, one places exactly that many dominoes (e.g., 641) end-to-end in a designated "input stretch" of the network. When the first domino tips, it initiates a cascade that includes all the dominoes in the input stretch. This triggers a series of processes throughout the network, including various loops that test the input number for divisibility by different potential factors (2, 3, 5, etc.).

If any of these tests finds a divisor, a signal travels down a specific path called the "divisor stretch," with falling dominoes indicating that the input number is not prime. Conversely, if all divisibility tests fail (meaning no divisors are found), a signal travels down a different path called the "prime stretch," with falling dominoes confirming the number's primality. The system thus physically implements the algorithm for primality testing through the propagation of falling dominoes. The physical arrangement of dominoes embodies the logical structure of the primality test, with each part of the network serving a specific computational purpose—whether testing divisibility by a particular number, processing the results of these tests, or signaling the final outcome.

B Bohr's Atomic Theory

Bohr's atomic theory (Bohr 1913), proposed by Niels Bohr in 1913, was developed to address a specific problem in physics: explaining the discrete spectral lines emitted by atoms. When elements are heated or subjected to electrical discharges, they emit light that forms a unique pattern of discrete lines rather than a continuous spectrum when passed through a prism. This phenomenon contradicted classical physics, which predicted that electrons orbiting a nucleus would emit a continuous spectrum of electromagnetic radiation.

To explain these observations, Bohr introduced several radical postulates. First, he proposed that electrons can only orbit the nucleus in certain discrete, stable orbits (energy levels) where they do not emit radiation. Second, he suggested that electrons can jump between these allowed orbits. When an electron moves from a higher-energy orbit to a lower-energy orbit, it emits a photon with energy equal to the difference between the two orbital energy levels. The frequency of this photon corresponds directly to a specific spectral line.

This mechanism provided a direct explanation for why spectral lines appear at precise frequencies rather than forming a continuous spectrum. Bohr's model successfully explained the observed hydrogen spectrum and introduced the concept of quantization to atomic physics, laying groundwork for the development of quantum mechanics.

C Euclid's Proof of Infinite Primes

Euclid's proof (Heath et al. 1956) that there are infinitely many prime numbers proceeds by contradiction. The proof begins by assuming that there are only finitely many prime numbers, which we can list as $p_1, p_2, p_3, \ldots, p_n$. Given this assumption, Euclid constructs a new number N by multiplying all these primes together and adding 1: $N = (p_1 \times p_2 \times p_3 \times \ldots \times p_n) + 1$

This number N is now examined. There are two possibilities: either N is prime, or N is composite (not prime). If N is prime, then we have found a prime number not in our original list, contradicting our assumption that we had listed all prime numbers. If N is composite, then N must be divisible by some prime number q. However, this prime q cannot be any of the primes in our original list $(p_1, p_2, \ldots, p_n)$ because dividing N by any of these primes always leaves a remainder of 1. Therefore, q must be a prime number not in our original list, again contradicting our assumption.

Since both cases lead to a contradiction, our initial assumption must be false. Therefore, there must be infinitely many prime numbers.

D Zagier's One-Sentence Proof

Zagier's original proof (Zagier 1990) that every prime $p \equiv 1 \pmod 4$ is a sum of two squares is famously presented in a single sentence. Here we provide a slightly more detailed exposition of the argument.

Theorem. Every prime $p \equiv 1 \pmod 4$ can be written as $p = a^2 + b^2$ for some integers a, b.

Proof. Consider the finite set $S = \{(x, y, z) \in \mathbb{N}^3 : x^2 + 4yz = p\}$. Since $x^2 \leq p - 4$ (as $y, z \geq 1$), we have $x \leq \sqrt{p-4}$, ensuring S is finite. Define an involution $f : S \to S$ by:

$$f(x,y,z) = \begin{cases} (x+2z, z, y-x-z) & \text{if } x < y - z \\ (2y - x, y, x - y + z) & \text{if } y - z < x < 2y \\ (x - 2y, x - y + z, y) & \text{if } x \geq 2y \end{cases}$$

This function is well-defined on S and satisfies $f(f(x, y, z)) = (x, y, z)$ for all $(x, y, z) \in S$. The involution f has exactly one fixed point: $(1, 1, (p-1)/4)$, where $p = 4k + 1$. By the *parity principle*, if an involution on a finite set has an odd number of fixed points, then the set itself has odd cardinality. Since f has exactly one fixed point, $|S|$ is odd.

Now consider the second involution $g : S \to S$ defined by $g(x, y, z) = (x, z, y)$, which simply swaps the y and z coordinates. Since $|S|$ is odd and g is an involution, g must have an odd number of fixed points by the parity principle. A fixed point of g satisfies $(x, z, y) = (x, y, z)$, which means $y = z$. Thus there exists $(x, y, y) \in S$ such that $x^2 + 4y^2 = p$. Setting $a = x$ and $b = 2y$, we obtain $p = a^2 + b^2$, completing the proof.

E Perspectives on World Models and Understanding

The following quotations illustrate a prevailing conception that links world models to understanding in artificial intelligence models.

LeCun (2022):

> *"Animals and humans exhibit learning abilities and understandings of the world that are far beyond the capabilities of current AI and machine learning (ML) systems. [...] By contrast, to be reliable, current ML systems need to be trained with very large numbers of trials [...]. Still, our best ML systems are still very far from matching human reliability in real-world tasks such as driving [...]. The answer may lie in the ability of humans and many animals to learn world models, internal models of how the world works."*

Ng (2023):

> *"Do large language models understand the world? [...] There's no widely agreed-upon, scientific test for whether a system really understands — as opposed to appearing to understand [...]. But with this caveat, I believe that LLMs build sufficiently complex models of the world that I feel comfortable saying that, to some extent, they do understand the world. To me, the work on Othello-GPT is a compelling demonstration that LLMs build world models; that is, they figure out what the world really is like rather than blindly parrot words."*

Mitchell (2025a):

> *"Of course, the word "understanding" is ill-defined, but one thing that seems key to human understanding is having mental "world models": compressed, simulatable models of how aspects of the world work, ones that capture causal structure and can yield predictions."*

Ser et al. (2025):

> *"A crucial shortfall of modern AI is the lack of World Models. Unlike humans, who construct internal representations to reason, predict, and decide, large models rely solely on statistical associations in their training data. A World Model enables AI to develop a structured, dynamic understanding of its environment, capturing relationships, rules, and causal links."*

Acknowledgements

We thank the anonymous reviewers and Kinshuk Vasisht for their valuable feedback. DP is grateful to Adobe Inc., Schmidt Sciences, Google and Microsoft Research for sponsoring his group's research.

References

Agarwal, N.; Ali, A.; Bala, M.; Balaji, Y.; Barker, E.; Cai, T.; Chattopadhyay, P.; Chen, Y.; Cui, Y.; Ding, Y.; Dworakowski, D.; Fan, J.; Fenzi, M.; Ferroni, F.; Fidler, S.; Fox, D.; Ge, S.; Ge, Y.; Gu, J.; Gururani, S.; He, E.; Huang, J.; Huffman, J.; Jannaty, P.; Jin, J.; Kim, S. W.; Klár, G.; Lam, G.; Lan, S.; Leal-Taixe, L.; Li, A.; Li, Z.; Lin, C.-H.; Lin, T.-Y.; Ling, H.; Liu, M.-Y.; Liu, X.; Luo, A.; Ma, Q.; Mao, H.; Mo, K.; Mousavian, A.; Nah, S.; Niverty, S.; Page, D.; Paschalidou, D.; Patel, Z.; Pavao, L.; Ramezanali, M.; Reda, F.; Ren, X.; Sabavat, V. R. N.; Schmerling, E.; Shi, S.; Stefaniak, B.; Tang, S.; Tchapmi, L.; Tredak, P.; Tseng, W.-C.; Varghese, J.; Wang, H.; Wang, H.; Wang, H.; Wang, T.-C.; Wei, F.; Wei, X.; Wu, J. Z.; Xu, J.; Yang, W.; Yen-Chen, L.; Zeng, X.; Zeng, Y.; Zhang, J.; Zhang, Q.; Zhang, Y.; Zhao, Q.; and Zolkowski, A. 2025. Cosmos World Foundation Model Platform for Physical AI. arXiv:2501.03575.

Andreas, J. 2024. Language Models, World Models, and Human Model-Building. Language & Intelligence @ MIT Blog.

Avigad, J. 2008. Understanding Proofs. In *The Philosophy of Mathematical Practice*. Oxford University Press. ISBN 9780199296453.

Baumberger, C.; Beisbart, C.; and Brun, G. 2016. What is understanding? An overview of recent debates in epistemology and philosophy of science. *Explaining understanding*, 1–34.

Bohr, N. 1913. I. On the constitution of atoms and molecules. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 26(151): 1–25.

Bruce, J.; Dennis, M.; Edwards, A.; Parker-Holder, J.; Shi, Y.; Hughes, E.; Lai, M.; Mavalankar, A.; Steigerwald, R.; Apps, C.; Aytar, Y.; Bechtle, S.; Behbahani, F.; Chan, S.; Heess, N.; Gonzalez, L.; Osindero, S.; Ozair, S.; Reed, S.; Zhang, J.; Zolna, K.; Clune, J.; de Freitas, N.; Singh, S.; and Rocktäschel, T. 2024. Genie: Generative Interactive Environments. arXiv:2402.15391.

Ding, J.; Zhang, Y.; Shang, Y.; Zhang, Y.; Zong, Z.; Feng, J.; Yuan, Y.; Su, H.; Li, N.; Sukiennik, N.; et al. 2024. Understanding World or Predicting Future? A Comprehensive Survey of World Models. *arXiv preprint arXiv:2411.14499*.

Friedman, M. 1974. Explanation and scientific understanding. *the Journal of Philosophy*, 71(1): 5–19.

Google DeepMind. 2025. Veo. https://deepmind.google/models/veo/. Accessed: November 2025.

Grimm, S. 2011. Understanding. In Berneker, D. P. S., ed., *The Routledge Companion to Epistemology*. Routledge.

Grimm, S. R. 2010. The goal of explanation. *Studies in History and Philosophy of Science Part A*, 41(4): 337–344.

Grimm, S. R. 2017. "Understanding and Transparency". In Baumberger, S. G. C.; and Ammon, S., eds., *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*. Routledge.

Hamami, Y.; and Morris, R. L. 2024. Understanding in mathematics: The case of mathematical proofs. *Noûs*, 58(4): 1073–1106.

Heath, T. L.; et al. 1956. *The thirteen books of Euclid's Elements*. Courier Corporation.

Hofstadter, D. R. 2007. *I am a strange loop*. Basic books.

Howard, W. A.; et al. 1980. The formulae-as-types notion of construction. *To HB Curry: essays on combinatory logic, lambda calculus and formalism*, 44: 479–490.

Hu, Z.; and Shu, T. 2023. Language models, agent models, and world models: The law for machine reasoning and planning. *arXiv preprint arXiv:2312.05230*.

Karai, K. 2010. Zagier's one-sentence proof of a theorem of Fermat. MathOverflow:https://mathoverflow.net/q/31113.

Karvonen, A. 2024. Emergent world models and latent variable estimation in chess-playing language models. *arXiv preprint arXiv:2403.15498*.

Karvonen, A.; et al. 2024. OthelloGPT Learned a Bag of Heuristics. LessWrong (blog post).

LeCun, Y. 2022. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1): 1–62.

Li, K.; Hopkins, A. K.; Bau, D.; Viégas, F.; Pfister, H.; and Wattenberg, M. 2023. Emergent world representations: Exploring a sequence model trained on a synthetic task. *ICLR*.

Mitchell, M. 2025a. LLMs and World Models, Part 1. AI Guide Substack.

Mitchell, M. 2025b. LLMs and World Models, Part 2. AI Guide Substack.

Ng, A. Y. 2023. Does AI Understand the World? Blog post in DeepLearning.AI's newsletter "The Batch".

Nikankin, Y.; Reusch, A.; Mueller, A.; and Belinkov, Y. 2024. Arithmetic without algorithms: Language models solve math with a bag of heuristics. *arXiv preprint arXiv:2410.21272*.

OpenAI. 2024. Sora: Creating video from text. https://openai.com/sora. Accessed: November 2025.

Poincaré, H. 1914. *Science and Method*. Mineola, N.Y.: Dover Publications.

Popper, K. R. 1945. *The Open Society and Its Enemies, Volume II*. Routledge. Chapter 11, Section II.

Popper, K. R. 1979. *Objective knowledge: An evolutionary approach*. Clarendon press Oxford.

Ser, J. D.; Lobo, J. L.; Müller, H.; and Holzinger, A. 2025. World Models in Artificial Intelligence: Sensing, Learning, and Reasoning Like a Child. arXiv:2503.15168.

World Labs. 2025. Generating worlds. https://www.worldlabs.ai/blog. Accessed: November 2025.

Xing, E.; Deng, M.; Hou, J.; and Hu, Z. 2025. Critiques of World Models. *arXiv preprint arXiv:2507.05169*.

Yang, D.; Hu, L.; Tian, Y.; Li, Z.; Kelly, C.; Yang, B.; Yang, C.; and Zou, Y. 2024. WorldGPT: a Sora-inspired video AI agent as Rich world models from text and image inputs. *arXiv preprint arXiv:2403.07944*.

Zagier, D. 1990. A one-sentence proof that every prime p$\equiv$ 1 (mod 4) is a sum of two squares. *The American Mathematical Monthly*, 97(2): 144–144.