Beyond World Models: Rethinking Understanding in AI Models

Anonymous ACL submission

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 position paper, we argue that human-level understanding extends beyond mental world models alone. We examine illuminating cases from prior philosophical work on understanding—including analyses of computational systems, physical theories, and mathematical proofs—to demonstrate how human understanding goes beyond just mental world models. By highlighting these distinctions, we hope to stimulate deeper discussion about what constitutes true understanding in both human and artificial contexts.

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). World models can thus be conceptualized as systems that track distinct states and the causal relationships between them, allowing predictions by maintaining representations of entities, their states, and the rules governing transitions (see §A for related work on world models in AI).

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. It is important to note that while people generally talk about mental models of the real world, we do not have direct access to the real world—we receive nerve signals from sense organs which are converted to electrical activity in the brain—hence, what we really mean is a world model of some theory of the real world, like the heliocentric model. Therefore, such mental models can exist even for superseded theories, like the geocentric model with its epicycles, or Bohr's model of electrons orbiting the atomic nucleus in discrete paths.

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 systems implies human-like understanding capabilities (LeCun, 2022; Ng, 2023; Mitchell, 2025a; Ser et al., 2025). **In this position paper, however, we argue that while world models represent an advance beyond mere surface patterns, they fail to capture human-level understanding across various domains of physical reasoning and problem-solving.**

While both world models and understanding lack

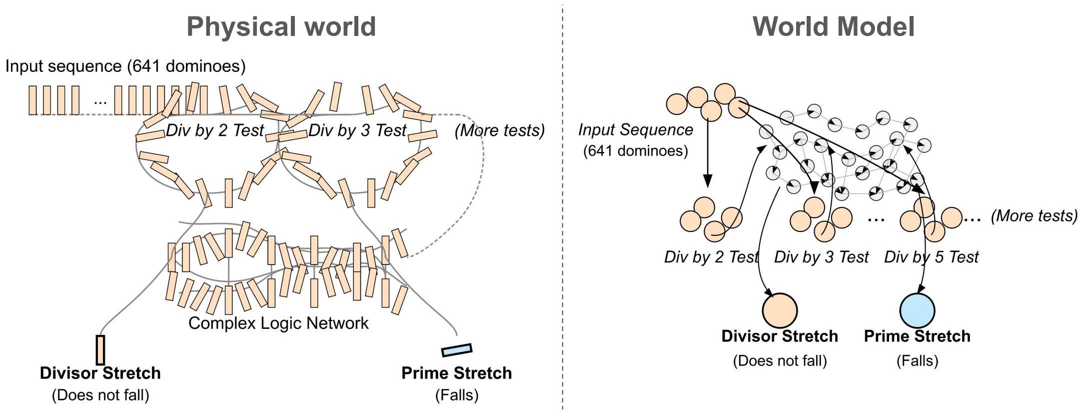

Figure 1: Left: Conceptual illustration of physical arrangement of dominoes in a computational system (Hofstadter, 2007). Right: A schematic world-model representation showing states and causal relationships between dominoes. While the world model can track physical states (standing or fallen dominoes) 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. (Note: Text labels on right diagram are included to improve the readability of the diagram, but are not a part of the actual state representation.)

universally agreed-upon definitions, the growing interest in world models within AI research makes exploring the distinction between world models and understanding valuable. We examine three cases from philosophical work (see §A for related work): (1) Hofstadter (2007)'s analysis of a computer built from falling dominoes, (2) Popper (1979)'s account of understanding physical theories through their problem situations, and (3) Pólya (1949)'s distinction between verifying and understanding mathematical proofs. These cases demonstrate that understanding exists in multiple degrees and the world-model conception falls significantly short of human-level understanding.

2 Case Studies: Understanding Beyond World Models

This section examines three cases where world models fall short of human-level understanding: a computer built from falling dominoes, Bohr's atomic theory, and mathematical proofs.

2.1 Understanding a Computer Built from Dominoes

Consider Hofstadter (2007)'s thought experiment of a computer built from millions of spring-loaded dominoes. In this system, 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 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. (For a more detailed explanation, see §B.)

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. Figure 1 provides a schematic illustration of this conceptual arrangement.

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 asked why a particular domino never falls when the input is 641, such a model might answer, "Because 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 the true understanding. The true 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.

2.2 Understanding Physical Theories

Popper (1979) argues that understanding a physical theory means understanding the problem situation that led to proposing that particular theory as a solution. By "problem situation," he means not only the problems one tries to solve but also their historical context—the problems together with theories that failed to solve them, and the criticisms that generated new problems requiring solutions. Understanding involves grasping this cycle of problem, tentative solution, criticism, and new problem—that is, why a theory failed and how a new problem emerged. It encompasses the entire historical situation surrounding the problem.

Consider Bohr's atomic theory (Bohr, 1913)[1]: Bohr proposed that electrons orbit the nucleus in discrete, fixed energy levels rather than in continuous paths as described by classical physics. The key to understanding it is not merely visualizing 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 truly 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 theory's purpose and significance. World models in AI similarly emphasize internal simulation—like picturing electrons on orbits—but as Popper (1979) argues, picturing is not understanding. An AI model might successfully simulate atomic transitions without grasping their importance in broader theoretical context, just as someone might visualize Bohr's orbital structure without comprehending its explanatory role in solving the spectral line problem.

2.3 Understanding Mathematical Proofs

Mathematical proofs, while often considered purely abstract logical structures, have a fascinating connection to computation through the Curry-Howard correspondence (Howard et al., 1980). This correspondence establishes an isomorphism between formal proofs and computer programs—every valid proof can be mapped to a computation that produces the conclusion from the premises, and every correct computation corresponds to a proof that the output follows from the input. This isomorphism allows us to analyze how world models might approach mathematical reasoning, where logical steps can be viewed analogously to states and deductive reasoning to state transitions.

When asked why a particular conclusion holds, such a model would trace backward through the chain of logical states. For example, consider Euclid's famous proof that there are infinitely many primes (Heath et al., 1956) (see §D for the proof). 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." The state before that might show "But this contradicts our earlier result that $N + 1$ is divisible by no prime on our list." 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

[1] For a quick refresher of Bohr's theory, see §C.

a proof and verifying it. As Poincare (1914) observes:

> *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? [...]*
>
> *[Majority of mathematicians] 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." This parallels our domino computer example, where tracking the sequence of physical states fails to reveal the abstract mathematical principles that explain why the system works. Pólya (1949) calls such a superficial understanding "deus ex machina" ("God from the machine"). Pólya (1949) further explains what constitutes better understanding:

> *Look here, I am not here just to admire you. I wish to learn how to do problems by myself. Yet I cannot see how it was humanly possible to hit up on your ... definition. So what can I learn here? How could I find such a ...definition by myself?*

As per this, truly understanding a proof involves re-enacting the discovery process. This requires comprehending, to some extent, the subject's history and the problems the author attempted to solve. Pólya (1949) terms this "a plausible story of discovery"—understanding the subject's history, the intermediate problems, their tentative solutions, why these solutions failed, and how they led to the final approach. We refer readers to Pólya (1949) for an excellent detailed walkthrough of this distinction applied to a real analysis problem.

This idea of "a plausible story of discovery" is remarkably similar to Popper (1979)'s "problem situation" concept discussed earlier. The independent development of similar ideas across different domains suggests this conception of understanding may capture something common to how humans understand across various fields.

3 A Possible Counterargument

A counterargument can be proposed that world models could include 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 a state and connected to the physical configurations of different dominoes. Similarly, for Bohr's theory, discrete spectral lines could be represented as a state and mapped to the mental picture of electrons orbiting in discrete paths.

However, if "states" in a world model can encode rich abstractions (e.g., mathematical properties, problem-solving strategies, historical context), then the concept of a world model becomes vacuously powerful—virtually anything can be labeled a world model if states are sufficiently enriched. This undermines the explanatory value of the concept itself. If the explanatory work is done by the state representation rather than the model's dynamics, calling it a "world model" adds little to our understanding of the agent's capabilities—it simply presupposes understanding in the state representations rather than explaining how it emerges.

4 Conclusion and Future Directions

In this position paper we have argued that mental world models, while superior to textual correlations, fall short of human-level understanding across our three case studies of the domino-based computer, Bohr's atomic theory, and mathematical proofs. For future directions, connections between philosophical theories of understanding and AI research could be productive. For example, Hamami and Morris (2024) have already formalized similar ideas to those we discuss in §2.3, proposing diagnostics for failures of understanding in mathematical proofs. Their framework might help develop behavioral benchmarks to assess how well LLMs understand proofs. Similarly, insights from philosophical theories of understanding (Popper, 1979) and explanations (Deutsch, 2011) (explanations provide understanding, the latter being the goal of the former) could inform future research directions. Our work advances the AI research discourse by highlighting the fundamental gap between world models and human-level understanding.

5 Limitations

While our work highlights an important distinction between human-level understanding and world-models, testing for the understanding outlined in this position paper in LLMs remains challenging. While prompt-based behavioral tests may be created for our case studies, these tests cannot reliably distinguish the understanding described in this paper from surface-level pattern recognition. Contrasting world-models with the understanding defined in our analysis using mechanistic interpretability methods remains even more elusive.

References

Jacob Andreas. 2024. Language models, world models, and human model-building. Language & Intelligence @ MIT Blog.

Christoph Baumberger, Claus Beisbart, and Georg Brun. 2016. What is understanding? an overview of recent debates in epistemology and philosophy of science. *Explaining understanding*, pages 1–34.

Vivien Beattie IV, Bill Collins, and Bill McInnes. 1997. Deep and surface learning: a simple or simplistic dichotomy? *Accounting education*, 6(1):1–12.

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

David Deutsch. 2011. *The beginning of infinity: Explanations that transform the world*. penguin uK.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. 2024. Understanding world or predicting future? a comprehensive survey of world models. *arXiv preprint arXiv:2411.14499*.

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

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. 2024. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*.

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

Yacin Hamami and Rebecca Lea Morris. 2024. Understanding in mathematics: The case of mathematical proofs. *Noûs*, 58(4):1073–1106.

Thomas Little Heath et al. 1956. *The thirteen books of Euclid's Elements*. Courier Corporation.

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

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

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

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

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

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent world representations: Exploring a sequence model trained on a synthetic task. *ICLR*.

Ference Marton and Roger Säljö. 1976. On qualitative differences in learning: I—outcome and process. *British journal of educational psychology*, 46(1):4–11.

Raphaël Millière and Cameron Buckner. 2024. A philosophical introduction to language models – part i: Continuity with classic debates. *Preprint*, arXiv:2401.03910.

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

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

Andrew Y. Ng. 2023. Does ai understand the world? Blog post in DeepLearning.AI's newsletter "The Batch".

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. 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`.

Andrés Páez. 2019. The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines*, 29(3):441–459.

Henri Poincare. 1914. *Science and Method*. Dover Publications, Mineola, N.Y.

George Pólya. 1949. With, or without, motivation? *The American Mathematical Monthly*, 56(10):684–691.

Karl Popper. 1979. *Objective knowledge: An evolutionary approach*, volume 49. Clarendon press Oxford.

Duncan Pritchard. 2014. Knowledge and understanding. In *Virtue epistemology naturalized: Bridges between virtue epistemology and philosophy of science*, pages 315–327. Springer.

Mirabel Reid and Santosh S Vempala. 2025. Does gpt really get it? a hierarchical scale to quantify human and ai's understanding of algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1492–1500.

Javier Del Ser, Jesus L. Lobo, Heimo Müller, and Andreas Holzinger. 2025. World models in artificial intelligence: Sensing, learning, and reasoning like a child. *Preprint*, arXiv:2503.15168.

Daniel A Wilkenfeld. 2013. Understanding as representation manipulability. *Synthese*, 190:997–1016.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. 2023. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pages 2226–2240. PMLR.

Deshun Yang, Luhui Hu, Yu Tian, Zihao Li, Chris Kelly, Bang Yang, Cindy Yang, and Yuexian Zou. 2024. Worldgpt: a sora-inspired video ai agent as rich world models from text and image inputs. *arXiv preprint arXiv:2403.07944*.

A Related Work

World Models in AI. The term "world model" has become a buzzword in the AI community and is considered by many a key ingredient for building general intelligence (LeCun, 2022; Ding et al., 2024). However, it does not have a universally accepted definition. Millière and Buckner (2024) describe it as internal representations that simulate aspects of the external world. LeCun (2022) describes it as a system that captures causal relationships between world states, enabling an agent to simulate outcomes, reason, and plan by representing entities and rules governing their transitions.

For LLMs, Li et al. (2023) arrive at a landmark result showing that a language model trained to play Othello developed an internal world model of the game. Similar evidence of internal world models has been found in LLMs trained on chess (Karvonen, 2024). Some research suggests these world models may not be clean, human-like mental models, but rather 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).

Recent advances in multimodal models have expanded world models to explicit simulators predicting future physical states. Models like Sora (OpenAI, 2024) and WorldGPT (Yang et al., 2024) function as world simulators by generating videos that aim to approximate physical laws. These models maintain temporal consistency while simulating physical interactions. Beyond videos, embodied world models create interactive environments for robotics (Wu et al., 2023) and autonomous driving (Gao et al., 2024).

For a comprehensive survey on world models, we refer readers to (Ding et al., 2024).

Philosophical Perspectives on Understanding. "Understanding" has been a subject of intense debate and scholarly investigation in philosophy literature for decades, with numerous competing theories and frameworks proposed to explain its nature (Popper, 1979; Pritchard, 2014; Baumberger et al., 2016; Páez, 2019). Explanation and understanding are closely related, where the latter is seen as the goal of the former (Friedman, 1974; Grimm, 2010; Deutsch, 2011). A key distinction in this literature is that understanding transcends mere knowledge acquisition. As Pritchard (2014) argues, the concepts of "knowing why" and "understanding why" may overlap but are not identical.

Wilkenfeld (2013) proposed the Understanding as Representation Manipulability (URM) framework, defining understanding as the ability to manipulate mental representations to produce useful inferences. This perspective aligns with educational theories distinguishing between "deep" and "surface" learning (Marton and Säljö, 1976; Beattie IV et al., 1997), where surface learning focuses on memorization while deep learning emphasizes conceptual connections. Recent work by Reid and Vempala (2025) has applied these philosophical frameworks to develop hierarchical scales for quantifying understanding of algorithms in both humans and AI systems.

Central to our analysis is Popper (1979)'s concept of "problem situation," which argues that understanding a theory means understanding the problems it was designed to solve, along with their historical context. A problem situation encompasses the entire historical cycle: problem, tentative solution, criticism, and new problem—understanding why a theory failed and how a new problem emerged. Poincare (1914) makes a similar observation about mathematical understanding, noting that it goes beyond examining each syllogism to grasping why they are linked in a particular order rather than appearing "engendered by caprice." Pólya (1949)'s "plausible story of discovery" complements these views—true understanding requires re-enacting the discovery process rather than merely verifying logical steps. These frameworks all highlight how understanding involves grasping historical context, problem-solving motivation, and the path of discovery.

B 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 feature allows signals to propagate through the system repeatedly, enabling complex computational processes. The domino chainium functions as a mechanical computer where signals travel through carefully arranged networks of dominoes. These signals can bifurcate (split into multiple paths), join together, and propagate in loops—creating a physical implementation of a computer program. The precise timing of domino falls is crucial to the functioning of this system, as it 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.

C 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.

D 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.