Why Does AI Lie So Much? The Problem Is More Deep Rooted Than You Think

Addressing The Semantic Grounding Problem of AI and How It Leads To Incurable Hallucinations

Mir H. S. Quadri | Arkinfo Notes | 5th August 2024

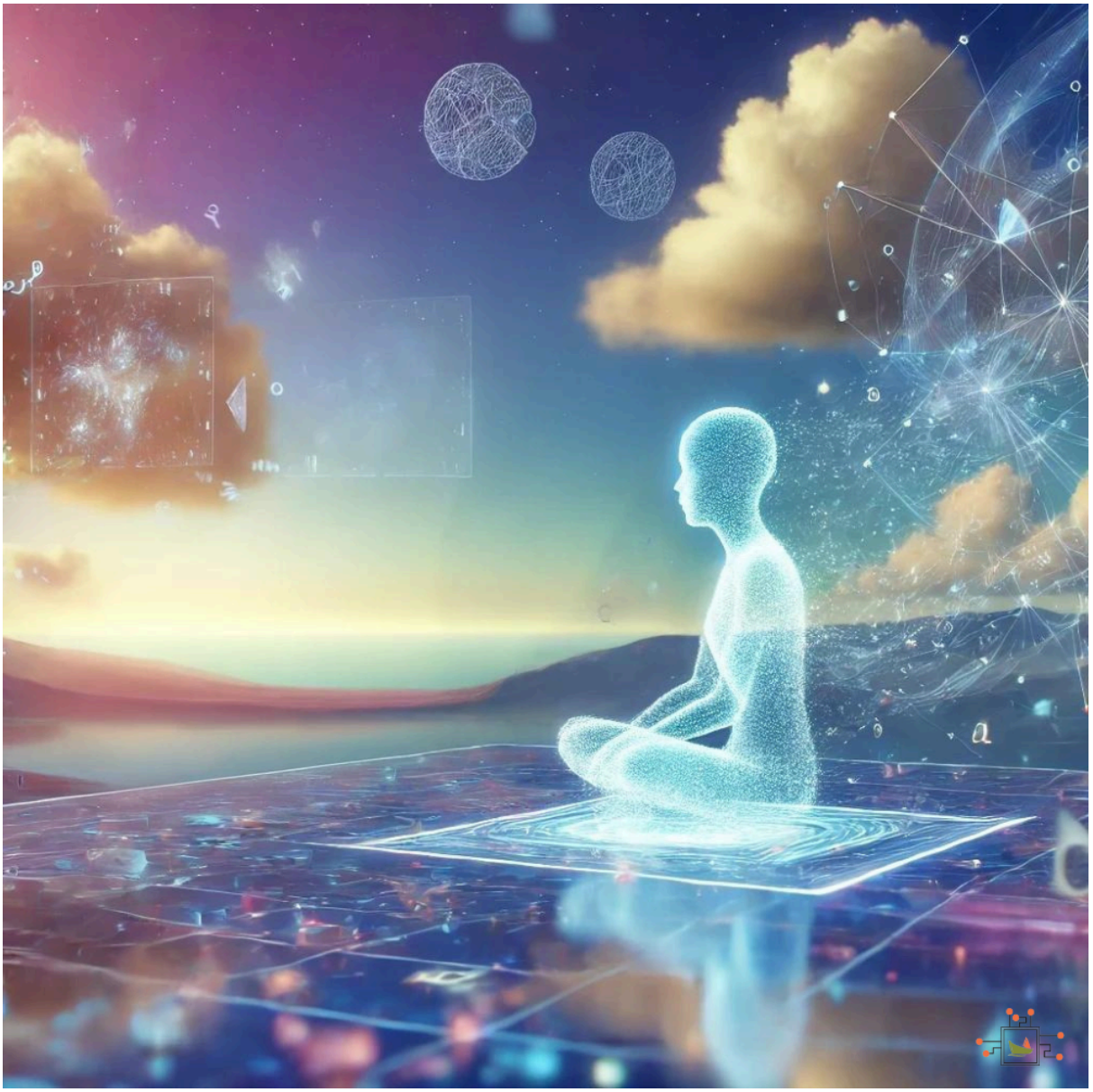
Let me start with a sentence containing 5 words.

Hulabalu hubla hubu haba hablo.

I made this sentence up. It doesn't belong to any language family. It consists of 5 words, i.e., *"hulabalu, hubla, hubu, haba, hablo"*. What do these words mean and what makes this sentence construction correct? What if I tell you that these 5 words mean nothing, literally or figuratively, and that the sentence construction is correct simply because I say it is? You'd probably think I am crazy. You'd be right. Why? Because words need to mean something. They need to represent an abstract notion, and also need to be placed in an order that brings *'sense'* to the notion that you are trying to convey.

But here's the thing. Suppose I created a billion such sentences, whose words have no meaning and whose structures are merely repetitive patterns without any inherent *sense* to them. Then I take those sentences and train a deep neural network on them, guess what will happen? The neural network will start to identify patterns in these words and sentences and start generating output that sounds exactly like the sentence that I shared above. AI can learn a completely *nonsensical* language that breaks all the rules of grammar known to man, and start generating content in that language.

The problem with such a form of *language learning* (if you can call it that) is that it is just mere surface level probabilistic pattern recognition. This is a far cry from what we as humans do with language.



Artwork - Semantic Grounding

As humans, one of the ways that we use language is to communicate abstract notions that have *semantic grounding,* i.e., share ideas that make sense to us. If I start calling the moon a *'cheese globe',* people will think me insane. No matter what I do, whether I write a book about it or dedicate my life to preaching that the moon is nothing but a globe of

cheese, no one is going to believe me because it makes no *sense*. In other words, it lacks *semantic grounding*.

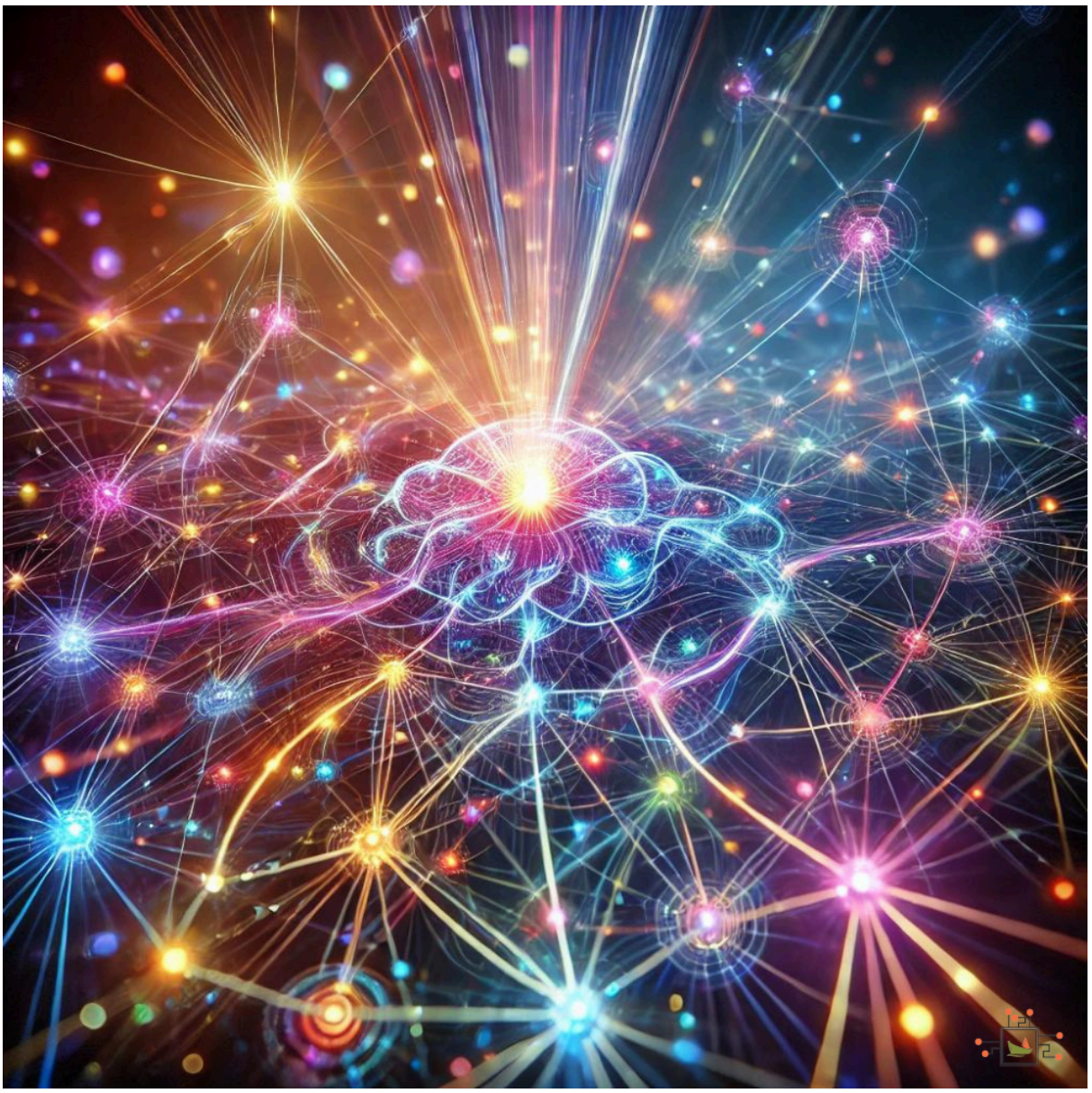
S*emantic grounding* is a phrase that I coined in an attempt to succinctly convey the lack of embodied cognition that AI has. In this article, I want to explore this concept deeply, trying to break down the flaws in our approach to AI development.

Our Obsession with Connectionism in AI

Let's start with the basics. Connectionism is a theoretical framework for understanding cognitive processes and mental phenomena through the lens of artificial neural networks. Imagine a network of neurons, not unlike the one in our brains, but vastly simplified and simulated on a computer. These artificial neurons, or nodes, are interconnected, and each connection has a *weight* that can be adjusted during learning. When you feed data into this network, it learns by tweaking these weights to minimise errors and improve accuracy.

Connectionist models, particularly deep neural networks, have achieved remarkable success in a variety of tasks. From recognising speech and images to translating languages and generating human-like text, these models have proven their prowess. They promise a future where machines can understand and generate language, drive cars, diagnose diseases, and maybe even surpass human intelligence. ChatGPT is a child of connectionism. No wonder the tech space is obsessed with connectionist philosophy. It gets the job done for the most part.

But as with any obsession, there's a downside. Despite their impressive capabilities, connectionist models are fundamentally limited. They excel at identifying patterns in data and making predictions based on those patterns. However, they do so without *truly understanding* the data. They are, in essence, statistical machines that *recognise correlations* rather than *comprehend meanings*.



Artwork - Connectionism

Let's bring this back to our initial example. If I create a billion *nonsensical* sentences and train a neural network on them, the network will undoubtedly learn to generate similar gibberish. It will become adept at mimicking the patterns it has seen, but it won't *understand* that the words mean nothing. This is the crux of the problem. Connectionist models operate on surface-level pattern recognition, lacking the deeper understanding that humans inherently possess.

This surface-level learning is the root cause of the hallucination problem. When AI models generate text, they rely on the patterns they've learned, but without an underlying structure or true comprehension, they can produce outputs that they deem to be grammatically correct yet semantically void or even factually incorrect. They can spin sentences that sound plausible but lack the grounding in reality that human language inherently has.

The tech field's obsession with connectionism has led to incredible advancements, no doubt. But it has also resulted in models that, while powerful, are fundamentally flawed in their understanding of language. They can predict the next word in a sentence but cannot grasp the meaning behind the words.

Connectionism has driven significant progress in AI. However, its reliance on pattern recognition without *deep understanding* is a critical flaw. This obsession with connectionism has led to the hallucination problem.

Why the Hallucination Problem is Significant

AI models, built on connectionist principles, are trained to identify and replicate patterns in data. They learn from vast datasets filled with text, images, and other forms of information, adjusting their internal weights to improve their performance. However, this learning is fundamentally shallow. It focuses on statistical correlations rather than understanding the underlying meaning or context of the data.

Consider the example of a language model generating text. When prompted, it predicts the next word based on the patterns it has learned from its training data. If the data contains a high frequency of certain words following others, the model will generate similar sequences. But it

does so without any comprehension of the content. It doesn't know that *"hulabalu"* and *"hubla"* are meaningless. It simply replicates the patterns it has seen.

This pattern-based approach works well for many tasks, but it breaks down when the model encounters situations where context and understanding are crucial. For example, when asked to generate a scientific explanation or provide legal advice, the lack of true understanding becomes apparent. The model might produce text that sounds authoritative but is riddled with errors or fabrications. It might even create entirely new *"facts"* that have no basis in reality.

The hallucination problem shows a fundamental flaw in current AI approaches. While these models can generate impressive and often useful outputs, their lack of true understanding and semantic grounding leads to significant errors. This is a significant problem because on the one hand, we have the temptation to keep pumping out new, more efficient models without worrying about their lack of understanding. On the other hand, there is always the risk of these models generating outputs that could potentially cause someone their life.

Chomsky's Critique of Connectionism

Of course I had to quote Chomsky in this article. After all, he is not only one of the most vocal critics of connectionism, but also, his theories have revolutionised our understanding of language, particularly through his concept of Universal Grammar (UG). Chomsky's critique of connectionism and AI's reliance on pattern recognition is rooted in his belief in an innate, structured foundation for language. To understand his objections, we must first explore what Universal Grammar is and why it matters. I have written extensively on UG and you can read it on the link I shared above. For now, I will provide a short summary for this article.

Universal Grammar is the idea that the ability to *acquire language* is hard-wired into the human brain. According to Chomsky, all human languages share a common underlying structure, a set of grammatical principles and rules that are innate to the human mind. This framework enables children to learn complex languages rapidly and efficiently, despite the often limited and imperfect linguistic input they receive, an argument known as the *"poverty of the stimulus."*

Chomsky posits that language learning is not merely a process of absorbing patterns from the environment but is guided by these intrinsic grammatical structures. This theory explains why children can generate and understand sentences they have never heard before and why all human languages, despite their diversity, exhibit deep structural similarities.

Chomsky's critique of connectionism, and by extension the neural network-based models dominating AI today, is based on the following key points.

Lack of Innate Structure

Connectionist models learn through exposure to vast amounts of data, identifying statistical patterns and correlations. But they lack the intrinsic grammatical structures that Chomsky argues are essential for true language understanding. Without these innate structures, AI models can only mimic surface-level patterns, leading to issues like hallucinations.

Surface-Level Learning

Chomsky contends that connectionist models operate at a superficial level, recognising patterns without understanding the underlying principles of language. This is in stark contrast to the human ability to

grasp deep grammatical rules and apply them creatively and correctly in novel situations.

Generative Capacity

The most amazing thing about human language is its generative nature, i.e., the ability to produce and comprehend an infinite number of sentences, including those never encountered before. Chomsky argues that this capacity arises from our innate grammatical framework, something that connectionist models, with their reliance on learned patterns, fundamentally lack.

Context and Meaning

Human language understanding is deeply contextual and meaning-driven. We do not just string words together based on probability, we use language to convey and comprehend complex ideas grounded in real-world experiences and shared knowledge. Connectionist models, however, often miss this depth, leading to outputs that may be contextually inappropriate or semantically hollow.

Semantic Grounding

Imagine you're at a park and see a child pointing at a tree while exclaiming, *"Tree!"*. The child isn't just identifying a pattern of shapes and colours, they're linking the word to a real-world object they've seen, touched, and maybe even climbed. This connection between words and experiences is what I call *semantic grounding*. It's the foundation of how humans understand and use language, and it's the crucial element that current AI models lack.

Semantic grounding refers to the process of linking language to real-world experiences and context. It's about more than just recognising patterns in data, or tagging words with images, as is the case with

multi-modal models. It involves understanding the meaning and relevance of those patterns in a way that is connected to the physical world and human experience. Here's why semantic grounding is so important.

Embodied Cognition

Humans experience the world through their senses and actions. When we learn a word like *"apple,"* it's grounded in our sensory experiences of seeing, touching, tasting, and smelling an apple. This multisensory grounding helps us understand the concept of an apple beyond its mere appearance or shape.

Our physical interactions with the world help us understand abstract concepts. For instance, we comprehend spatial language *(e.g., "over," "under")* through our physical experiences of moving and navigating space.

Contextual Understanding

We learn words and their meanings through interactions with others, understanding not just what words mean, but how and when they are used. The meaning of words can change depending on the context. For example, the word *"bank"* can refer to a financial institution or the side of a river, depending on the context. Humans use situational cues to disambiguate such meanings effortlessly.

Cognitive Frameworks

Humans organise knowledge through cognitive structures like schemas and mental models. These frameworks help us make sense of new information by relating it to what we already know. When we encounter a new concept, we integrate it into our existing knowledge base, grounding it in our prior experiences and understanding.

Memory and Learning

Human memory systems form associations between words and their meanings based on repeated exposure and use in context. This associative memory allows us to retrieve and use words appropriately in various situations.

Humans have powerful learning mechanisms that enable us to extract patterns and regularities from our environment. We don't just memorise words, we understand their meanings and relationships through a process of active learning and contextual integration.

Real-World Knowledge

Our understanding of language is grounded in a rich network of real-world knowledge and experiences. We know that the moon is not made of cheese because of our scientific knowledge and observations, not just because it doesn't fit into a learned pattern.

Humans have the ability to verify and reason about information. If someone tells us the moon is made of cheese, we can draw on our knowledge and reasoning skills to challenge and refute that claim.

No, Multimodal Models are Not the Answer

Now before I close this article, it's important to address one key point, which is multimodal models. That's the next step the AI industry has taken and many see it as the answer to the semantic grounding problem. I beg to differ. Here's why.

Surface-Level Integration

Multi-modal models often excel at recognising patterns across different data types but still lack true understanding. For example, a model trained

on both images and text of apples can identify and generate descriptions of apples, but it does so based on statistical correlations rather than an inherent comprehension of what an apple is.

While these models can leverage multiple data sources, they often fail to integrate these sources in a way that captures the deeper contextual and experiential knowledge humans use. They might recognise that an image and a description match, but they don't truly grasp the sensory and functional experiences associated with the object.

Lack of Embodied Cognition

Multi-modal models still lack the ability to physically interact with the world. Human understanding is deeply rooted in embodied experiences, how we manipulate objects, move through space, and engage with our environment. AI models that only process sensory data without physical interaction miss a crucial component of semantic grounding.

Simulated experiences, such as watching a video of an apple, are not the same as real, tactile experiences. Humans use their entire sensory and motor systems to ground their understanding of concepts, a depth of engagement that current multi-modal models cannot replicate.

Static Knowledge Representation

Human understanding is dynamic, constantly updated through new experiences and interactions. AI models need a mechanism to continually integrate new knowledge and experiences to maintain relevance and accuracy. While multi-modal models can perform well within the confines of their training data, they often struggle to generalise beyond it. They may fail to apply their learned knowledge to new, unencountered scenarios in the same flexible and adaptive way humans can.

Cognitive and Contextual Disconnect

Human cognition involves complex mental models and schemas that help us understand and predict the world. These cognitive frameworks are built over time through rich, layered experiences. Multi-modal models, even with diverse data inputs, lack the depth and complexity of these human cognitive structures.

Without a deep, embodied understanding, multi-modal models can misinterpret context. They might link data points that appear related but miss the subtleties and nuances that human cognition naturally handles. This can lead to errors in judgement and comprehension, similar to the hallucination problem seen in purely text-based models.

A Course Correction is Needed

It is becoming increasingly clear that we face significant challenges in achieving true semantic grounding. Despite the remarkable advancements in AI, from chatbots to autonomous vehicles, the fundamental issue of understanding and contextualising language remains unresolved. The hallucination problem is a stark reminder that, without a solid grounding in real-world experiences and context, AI models will continue to generate outputs that, while plausible on the surface, lack the depth and reliability we expect.

Our current trajectory in AI development, heavily reliant on connectionism and pattern recognition, has brought us far. Yet, it is also evident that this approach has its limitations. The reliance on statistical correlations and vast datasets, without an underlying comprehension of meaning, has resulted in systems that can mimic human language but not truly understand it. This gap between mimicry and understanding is at the heart of the semantic grounding problem.

To address this, we must consider a course correction in our approach to AI research and development.

We must explore ways to enable AI systems to interact with the physical world in meaningful ways. This could involve developing robots or virtual agents that can engage with their environments, learning through direct experience rather than static data. Incorporating more diverse sensory inputs, beyond just visual and textual data, can help create a more nuanced understanding of concepts.

AI systems should be designed to learn and adapt continuously, integrating new experiences and information in real-time. Enhancing AI's ability to understand and apply context is essential. A hybrid approach that integrates the strengths of connectionist models with symbolic AI and cognitive frameworks could be a way to proceed. By combining pattern recognition with rule-based systems and mental models, we can create AI that is both flexible and grounded.

Developing AI systems that can build and utilise cognitive frameworks similar to human schemas and mental models can enhance their ability to understand and generate meaningful language. Addressing biases in training data and ensuring diverse and representative datasets is also needed.

Building AI systems that are transparent in their operations and decisions will create accountability and trust. Can blockchain be used here? Just something to consider. Users should be able to understand and challenge AI outputs, especially in critical applications.

While we have made significant strides in AI development, the semantic grounding problem requires a fundamental shift in our approach. By integrating embodied cognition, dynamic and contextual learning, hybrid models, and committing to ethical development practices, we can

pave the way toward AI systems that truly understand and interact with the world in a meaningful way.
