A Primer on Knowledge And Where It Comes From

The Machine Mind: From Blank Slate to World Models – How AI is Redefining "Knowledge"

The quest to build artificial intelligence (AI) is not merely a technical challenge; it is a profound philosophical experiment that forces us to confront fundamental questions about knowledge itself. For centuries, philosophers have grappled with how humans acquire understanding. Today, the debate rages anew, not in the ivory towers of academia, but in the silicon valleys of AI research, where different approaches to AI reflect contrasting philosophical traditions. By examining the work of historical figures like John Locke and John Stuart Mill alongside contemporary AI leaders like Yann LeCun, we can illuminate how modern AI is, in essence, a grand, live-action reenactment of the problem of knowledge.

At the heart of this discussion lies **Empiricism**, a philosophical school championed by **John Locke** in the 17th century.[1] Locke famously argued for the concept of ***tabula rasa***, the "blank slate."[2] He contended that the human mind is not born with innate ideas, but rather acquires all knowledge through **experience**.[3] This experience comes in two forms: **Sensation** (our direct perception of the external world through our senses) and **Reflection** (the mind's observation of its own operations, like thinking, believing, and willing). For Locke, "raw data"—the myriad sights, sounds, textures, and smells—are the fundamental building blocks.[4] Our minds then process these "simple ideas" to form "complex ideas," recognizing patterns of similarities and differences, and crucially, observing **cause and effect**. We learn that fire burns, that a thrown stone falls, not because we are born knowing these truths, but because we constantly observe these phenomena in the world around us. John Stuart Mill later expanded on this, emphasizing **inductive reasoning**—the process of deriving general principles from specific observations.

The Spark of Reason—Locke vs. Hobbes

To understand why "embodiment" is the holy grail for modern AI, we must first understand a 17th-century schism regarding the human mind.

1. Hobbes and the "Calculating Machine"

Before Locke, **Thomas Hobbes** proposed a view of the mind that looks remarkably like a precursor to modern computer science. In *Leviathan*, Hobbes famously wrote, *"Reason is nothing but Reckoning."* To Hobbes, thinking was a form of addition and subtraction of sequences.[1] If you see a cloud, then you see rain, your mind simply adds those two "inputs" together.

In this view, the mind is a **passive processor**. It doesn't need to understand *why* things happen; it just needs to calculate the statistical likelihood of one thing following another. This is, in essence, the "Large Language Model" (LLM) view of the world: intelligence is the ability to calculate the next most likely word in a sequence based on a massive database of past "reckonings."

2. Locke: Reason as Agency and "Power"

John Locke disagreed. For Locke, reason was not just a calculator; it was the core of **human identity**. He argued that we don't just notice that B follows A; we develop the idea of **"Active Power."**[2]

Locke observed that when we see a billiard ball move another, we see an effect. But when we *decide* to move our own arm, we experience the "Power" of our will over the physical world. This is the "Aha!" moment of human reasoning. For Locke, reason is the ability to:

1. **Stop and Reflect:** We can pause our impulses to examine the cause-and-effect "laws" of our environment.[3]
2. **Internalize the "Why":** We don't just calculate sequences; we build a "mental model" of the forces (the "Powers") that make the world work.

For Locke, a human is not just a biological calculator (Hobbes); a human is an **agent** who understands their own place in the causal chain of the universe. This understanding is what allows us to take responsibility for our actions—it is the foundation of the "self."

If Hobbes is right, we might eventually build "Super-Intelligence" just by making bigger calculators (like GPT-5 or 6). But if Locke is right, we will never achieve true intelligence until we give the "blank slate" a pair of eyes, a set of hands, and the ability to drop a glass and watch it shatter.

True reasoning isn't just knowing what happens next; it's understanding the "Power" that makes it happen.

Now, let's fast-forward to the 21st century and the landscape of AI. The dominant form of AI that has captivated the public imagination in recent years is the **Large Language Model (LLM)**, exemplified by systems like ChatGPT.[5] LLMs are trained on vast datasets of human-generated text and code—the entirety of the internet, books, articles, and more.[6] Their brilliance lies in their ability to predict the next word in a sequence, generating remarkably coherent and contextually relevant text.

From an empiricist perspective, however, LLMs present a fascinating paradox. They demonstrate an extraordinary command of what we might call "second-hand knowledge." An LLM has "read" virtually everything humanity has ever written. It understands the patterns of human language, reasoning, and even bias, as encoded in text. It can synthesize, summarize, and even generate creative content based on these linguistic patterns. In Lockean terms, an LLM operates almost exclusively on the "Reflection" aspect of knowledge, but only as it pertains to human *linguistic* reflection. It knows *what we say* about the world, but not the world itself. It understands the map, but has never journeyed through the territory. It possesses knowledge that has been thoroughly **"pre-digested"** by human perception, categorized, labeled, and encoded into symbolic language.

This limitation is precisely what AI pioneers like **Yann LeCun** (Chief AI Scientist at Meta) and Fei-Fei Li (Stanford University) are critiquing. They argue that an AI trained solely on text can never truly "know" the world in the way a human or even an animal does. It lacks **common sense** because it has never experienced the fundamental physical laws governing our reality. It has no intrinsic understanding that objects fall, that a cup is opaque, or that pushing a block makes it move.

This critique leads us directly to the concept of **Embodied AI** and LeCun's proposed solution: the **Joint Embedding Predictive Architecture (JEPA)**. Imagine a baby observing the world. It doesn't need someone to label every object or action; it simply watches, plays, and predicts. If it drops a toy, it anticipates it hitting the floor. If it sees a ball roll behind a couch, it knows the ball still exists, even out of sight. This is learning through raw **Sensation** and its own internal **Reflection** on those sensations to build a "world model."

JEPA aims to emulate this. Instead of training an AI to predict every pixel in a video (which is like predicting every single leaf fluttering in the wind—too much irrelevant detail), JEPA learns to predict the **abstract representations** or "embeddings" of future video frames.[7] The AI takes a raw sensory input (like a video frame) and compresses it into a high-level, semantic understanding in a "latent space."[8] It then tries to predict the abstract representation of what comes next.

This is a crucial departure:

- **From "Pre-Digested" to "Raw":** JEPA doesn't require human labels.[9] It learns from pure, unfiltered sensory data, just as Locke's *tabula rasa* gathers sensations.
- **From "Pattern-Matching" to "World-Modeling":** Rather than merely correlating words in text, JEPA builds an internal model of how the physical world works.[10] It learns "intuitive physics"—the cause-and-effect relationships that govern objects in space and time. It learns that objects are solid, that gravity exists, and that actions have consequences, not by being told, but by observing and predicting.
- **From "Human-Centric" to "World-Centric":** An embodied JEPA system could theoretically perceive and identify patterns in the world that humans cannot—perhaps in infrared light, ultrasonic frequencies, or subtle material stresses—and develop "categories" of knowledge that entirely escape human linguistic description. Its "knowledge" would be grounded not in human consensus, but in the objective regularities of the universe.

In essence, LeCun's JEPA is a modern attempt to build an AI that performs the Lockean act of acquiring knowledge from first principles. It bypasses the human filter, aiming to create an intelligence that actually "knows" something because it has done the fundamental work of experiencing, abstracting, and inferring patterns from the raw fabric of reality.

The ongoing debate between the prowess of LLMs and the promise of embodied AI like JEPA is, therefore, a contemporary re-framing of the age-old philosophical question: What truly constitutes knowledge? Is it the mastery of human-created symbols and narratives, or is it a direct, empirical understanding of the physical world itself? The answers we find in building AI will not only shape our technology but will continue to redefine our understanding of our own minds and our place in the universe.

The Bridge to LeCun: Why AI Must "Feel" Gravity

This brings us directly to the "Embodied AI" debate. Yann LeCun's argument is that LLMs are currently stuck in the **Hobbesian trap**: they are brilliant "reckoners" of text, but they have no "Lockean" sense of active power.

- **The LLM as Hobbesian:** An LLM knows that the word "gravity" is usually followed by the word "down." It has "calculated" this sequence millions of times. But it doesn't *understand* gravity because it has never felt the "power" of an object's weight in its own hand.
- **The Embodied AI as Lockean:** LeCun's JEPA (Joint Embedding Predictive Architecture) aims to move AI from "reckoning" to "reasoning."[4] By giving an AI a body (or a 3D simulation), the AI can finally experience **cause and effect** first-hand.

When a robot pushes a block and it falls, the AI isn't just predicting the next pixel; it is experiencing the **Active Power** Locke described. It is learning that *it* is an agent that can cause effects in the world. LeCun argues that this is the only way to "kick-start" true reasoning. Without independent experience of physics, an AI can never have "common sense" because it doesn't know what it's like to be a "cause" in a world of "effects."

Questions For Discussion

1. **The Sovereignty of Experience:** If Locke is right that knowledge requires first-hand sensation, does that mean an LLM (like the one you are talking to right now) is technically "ignorant" despite knowing every fact in the Library of Congress?
2. **The "Alien" Perspective:** If an embodied AI develops its own patterns from raw data and those patterns don't align with human language, is it still "knowledge" if it's not "social"? This hits on the social theory of **communicative action**.
3. **The Power Gap:** Locke's idea of "Active Power" is about the *Will*. Current AI has "objectives" (math), but humans have "desires" (biology). Can an AI ever truly "reason" if it doesn't have the biological stakes of survival that Locke assumed were part of the human experience?

Term	**Philosophical Context (Locke/Hobbes)**	**AI Context (LeCun/Li)**
Tabula Rasa	The "blank slate"; the mind starts without innate ideas.	**Self-Supervised Learning**: An AI that starts without human labels/tags.
Sensation	Raw data from the five senses (sight, touch, etc.).	**Raw Input**: Unfiltered data from cameras, LiDAR, and touch sensors.
Reflection	The mind's ability to think about its own patterns and ideas.	**Latent Space**: The internal mathematical "mental model" the AI creates.
Reckoning	Hobbes's idea that thinking is just calculation/addition.	**Predictive Text**: The mechanism of LLMs (statistical probability).
Active Power	The experience of being a "cause" in the world.	**Agency/Robotics**: The ability of an AI to take an action and observe the result.
World Model	A mental map of how reality works (gravity, physics).	**JEPA**: An architecture that predicts physical outcomes in a simulation.
Common Sense	Basic understanding of the world's physical "laws."	**Grounding**: Connecting abstract symbols (words) to physical reality.