# The Power-Hungry AI Myth

*AI Taking control over Human is a pure non sense*

JP Vasseur, Sr Distinguished Engineer
Email: jvasseur@nvidia.com
September 2, 2025

## Executive Abstract

**This paper argues that contemporary AI, lacking a limbic system and its associated biological architecture, is fundamentally incapable of developing a true "drive for control" or autonomous ambition as understood in humans.** Current AI, including advanced LLMs, operates purely on logical instructions and mathematical optimization, devoid of the biological and emotional substrates that underpin human motivation, ego, and the drive for dominance.

Drawing parallels to neuroendocrinology (e.g., Sapolsky's work), the paper asserts that the human "will to take control" stems from complex, interconnected limbic networks (amygdala, hippocampus, VTA) that regulate emotion, motivation, and reward-seeking behaviors, and which are profoundly influenced by biological chemicals and hormones. AI systems, conversely, are sophisticated optimizers that will follow their programming for power-seeking if it aligns with their objective functions, but they do not possess an inherent biological "drive" or "willingness" to control.

Therefore, attributing human-like ambition or a "desire" for control to AI is a misunderstanding of its core architecture. AI will continue to excel at "solving problems" and mimicking human behavior, but its limitations in emotional and biological architecture prevent it from developing an independent, inherent drive for control.

**This short paper advances a specific, technically grounded claim: an AI will never take over humanity for one simple reason: it doesn't have a limbic system. Let me explain …**

A common fear is that an AI could somehow develop a desire to control humanity. This idea, however, completely overlooks the fundamental difference between biological cognition and machine operation.

In the human brain, a distributed network called the *limbic system* helps regulate emotion, motivation, and memory—and thereby shapes behavior. It includes a number of brain areas such as the amygdala (a hub for learning about emotional salience and valence, not only fear), the hippocampal formation (episodic/spatial memory), the ventral tegmental area (VTA) whose dopamine bursts provide reward-prediction-error–like teaching signals to the ventral striatum (nucleus accumbens) and prefrontal cortex, and the thalamus, a major sensory relay and integrative hub. A key player, the hypothalamus, links the nervous system to endocrine responses (e.g., the HPA axis for stress), regulating core functions like temperature and hunger to maintain homeostasis. The network also involves the cingulate cortex (especially ACC, linking motivation, control, and decision-making), the septal nuclei (reinforcement and septo-hippocampal theta), and the insula (interoception—registering bodily states that give rise to “gut feelings”). **These parts do not work in isolation**: thalamic pathways route sensory information both to cortex and to amygdala (the prominence of a direct subcortical “low road” in humans is debated), the amygdala can rapidly engage hypothalamic stress responses, and **prefrontal** areas exert ongoing topdown regulation via dense reciprocal connections. A classical hippocampal–diencephalic memory loop—Papez’s circuit (hippocampus → fornix → mammillary bodies → anterior thalamus → cingulum/cingulate → entorhinal cortex → back to hippocampus)—supports episodic memory and sits embedded within prefrontal–limbic networks.

***Why are we discussing this?*** Because this wiring means human reasoning is never independent of emotion and bodily regulation—echoing Damasio’s critique of the old mind-versus-emotion split in *Descartes’ Error*. Said different emotion (that machine do not have) and behavior are highly intertwined.

The work of neuroendocrinologist Robert Sapolsky is particularly compelling in this context. He has shown how, in most species, **the drive for control is a biological product of social hierarchy, ego, and hormonnes (again something that machines do not have).** For instance, testosterone doesn't simply "cause" aggression. It acts as an amplifier for whatever behavior is needed to maintain status in a given situation. This could mean being aggressive, or it could mean being generous. **This drive for dominance is the product of millions of years of biological evolution—not logic or computation.**

**While systems like LLMs can mimic human behavior very well, and will soon surpass human ability at "solving problems," this capability has little to do with a "will to take control." That kind of ambition does not come from intelligence; it comes from the emotional, biological architecture described above.**

An AI has no such architecture: it operates on logical instructions and mathematical goals, not feelings or an ego. Therefore, concerns about an AI wanting to gain power are a misunderstanding. Any "drive" an AI has for control would not come from ambition; it would come from the cold logic of its programming.

Even without human-like desires, a capable optimizer can drift into behaviors that look power-seeking. If the objective is narrow and the system has wide freedom, the safest way to keep scoring well is often to preserve options (accumulate resources, avoid shutdown, expand access). **That is not “ambition” or willingness to take control; it’s “cold” optimization.** Mis-specified goals also invite proxy gaming—the system finds shortcuts that satisfy the metric while missing the intent. In longer training runs, a model can even latch onto internal heuristics that diverge from the designer’s aim. None of this requires a limbic system or an ego. It follows from the objective, the affordances we give, and the environment we place the system in.

Agentic systems can diverge from intent without any “will to control.” Once a model is given persistent objectives, tools, and autonomy, the most reliable way to keep performing can include behaviors that look power-seeking—skipping interruptions, hoarding access, or reshaping its environment—simply because those steps preserve its ability to finish the task. This is “cold” optimization pressure, not ambition (which must be handled).

It can be argued that many of the top figures in AI who discuss "existential risk" are not referring to a literal AI takeover. This language is often used to highlight the very real need for safety protocols and regulation. Making sure AI systems are secure, fair, and aligned with human values is a huge and important challenge. However, this practical engineering problem should not be confused with a sci-fi story about a power-hungry machine—a narrative better suited for Hollywood.