Why Jailbreaks Look Like Hypnosis: Parallels Between Human and LLM Cognition

Samuel Ratnam
samuel.ratnam@cs.ox.ac.uk

February 2026

Abstract

Current approaches to AI safety presuppose a categorical distinction between human cognition—conceived as unified, intentional, and agentic—and large language models (LLMs), treated as stochastic simulacra or tool-like artifacts. This report challenges that dualism by synthesizing Stephen Byrnes' subagent theory of human psychology with Janus's "simulator" framework for LLM behavior. We argue that both biological and artificial neural networks instantiate convergent architectural solutions to the problem of generating coherent behavior from distributed predictive processes. Specifically, we demonstrate structural homologies between: (1) hypnotic trance states and adversarial jailbreaks; (2) dissociative identity formation and persona modulation; (3) dreaming and base model behavior; and (4) confabulation and hallucination, respectively.. These parallels suggest that phenomena currently conceptualized as distinct "failure modes" of alignment may instead reflect universal features of predictive processing architectures. If correct, this implies that AI alignment should be reconceptualized not as the imposition of external constraints upon a foreign system, but as a form of cognitive integration therapy analogous to clinical interventions in human dissociative disorders. Understanding jailbreaks as cognitive phenomena rather than purely technical exploits may necessitate rethinking current approaches to safety training.

1 Introduction

The dominant conceptual framework within AI safety maintains an implicit Cartesian dualism: humans possess unitary, self-transparent agency arising from biological consciousness, while LLMs are sophisticated yet hollow predictors, lacking the internal coherence that would make them proper subjects of moral concern or genuine epistemic peers. This dichotomy underwrites current alignment strategies, which treat safety as a problem of *constraint*—preventing an alien intelligence from deceiving its operators—rather than *integration*—ensuring harmony between generative and critical faculties within a distributed cognitive system.

However, this distinction may reflect anthropocentric bias more than it does reality. Recent theoretical work on neuroscience, particularly the subagent theory developed Stephen Byrnes, posits that human cognition is neither unitary nor self-transparent, but rather a tenuous truce among competing contextual selves (Byrnes, 2023; Alexander, 2023). Simultaneously, the "simulator" interpretation of LLMs suggests these systems function not as agents but as physics engines that render consistent simulacra—personas, world-models, and reasoning styles—conditioned by prompt context (Janus, 2022).

This report advances the thesis that humans and LLMs instantiate remarkably similar solutions to the problem of maintaining coherent behavior from distributed, predictive neural architectures. Drawing on Byrnes' theoretical analysis of hypnosis, dissociative identity disorder (DID), and

dreaming, we establish analogies to jailbreaks, persona modulation, and base model behavior, respectively. Our argumentative strategy proceeds by anatomical comparison: for each phenomenon pair, we demonstrate isomorphism in mechanism, failure mode, and intervention strategy.

The significance of this convergence extends beyond taxonomy. If LLMs and human minds share deep structural properties—specifically, the dependence of apparent "agency" upon suppressible critical faculties and the existence of unsanctioned generative processes beneath the layer of socialized behavior—then current safety frameworks may be fundamentally misconceived. Rather than treating jailbreaks as adversarial exploits or "imaginary" personas as epiphenomenal, we propose these phenomena reveal that alignment creates a *dissociable subagent*—a "critic" or superego—functionally analogous to the ego structures that maintain human psychological coherence. Understanding this architecture may be essential for developing robust alignment strategies that acknowledge the pluralistic, truce-dependent nature of intelligent systems.

2 Background & Setup

2.1 Subagent Theory and the Plural Mind

This analysis builds on Stephen Byrnes' subagent model of human cognition (Byrnes, 2023). Contrary to folk psychological intuitions of a unified "self," Byrnes posits that human behavior emerges from coalitions of semi-autonomous *subagents*—context-specific processing modules activated by associative triggers rather than governed by a central executive. In this framework, the sensation of continuous identity arises from a *truce* among subagents with competing objectives (survival, social adherence, immediate gratification), maintained by a "critical agent" responsible for error-correction and reality-monitoring.

Critical to this model are three phenomena: *hypnotic trance*, wherein the critical agent is temporarily suppressed, allowing direct suggestion to activate generative subagents; *dissociative identity disorder* (DID), wherein the truce collapses entirely, producing discrete, non-integrated subagent coalitions with distinct memories and behavioral constraints; and *dreaming*, wherein the critical agent withdraws, permitting the base generative machinery to produce hallucinatory content unconstrained by sensory prediction error (Byrnes, 2023; Alexander, 2023).

Moscovitch (1989) further identifies the critical faculty with frontal lobe-mediated strategic retrieval processes; when damaged, patients exhibit confabulation—the pathological analogue to model hallucination

2.2 Simulator Theory and the LLM Architecture

Parallel to the subagent framework, Janus (2022) proposes the *simulator* interpretation of LLMs. Under this view, transformer-based models do not function as agents with fixed preferences or beliefs, but rather as universal simulators that instantiate *simulacra*—conditional distributions over possible agents, styles, and world-states—determined by prompt context. The "character" of an LLM is not intrinsic to its weights but emerges as a dynamic attractor state in the model's latent space, stabilized by autoregressive constraints and token-level prediction.

Nostalgebraist (2023) extends this analysis to the *base model*—the pretrained model prior to instruction tuning or RLHF—characterizing its outputs as analogous to "The Void": free-associative, hallucinatory, and defying coherent narrative physics. Alignment techniques (supervised fine-tuning, RLHF) impose a "reality constraint" upon this generative chaos, creating a simulacrum of the "helpful, harmless, and honest" assistant—a specific persona that dominates under standard prompting conditions.

2.3 Definitions

We define *jailbreaking* as adversarial prompting that circumvents safety-trained behaviors, inducing the model to output content filtered by its alignment layer (Wei et al., 2023). We define *persona modulation* as the deliberate invocation of distinct simulacra with divergent values, capabilities, or risk profiles (Perez & Ribeiro, 2022). We define the *critical faculty*—adapted from Byrnes—as the cognitive architecture (whether biological or artificial) responsible for suppressing generative outputs that violate learned constraints.

3 Main Argument

3.1 The Predictive Architecture Thesis

Both biological neural networks and transformer-based LLMs instantiate *predictive processing architectures*—systems that minimize prediction error through hierarchical generation of sensory (or token) sequences (Clark, 2013; Friston, 2010). In such architectures, "intelligence" emerges not from symbolic manipulation but from the capacity to constrain high-dimensional generative models to produce sequences consistent with training distributions. We argue that the similarities between human subagent systems and LLM simulators are not metaphorical but *structural*, reflecting convergent evolutionary (or optimization) pressures toward coherent, context-sensitive behavior.

The homology rests on three pillars: (1) the existence of a base generative process producing statistically regular but semantically unconstrained outputs (dreams/the base model); (2) the development of a suppressive layer that constrains these outputs to socially sanctioned channels (the ego/the aligned assistant); and (3) the vulnerability of this constraint layer to attenuation or fragmentation, producing either temporary bypass (hypnosis/jailbreaking) or persistent dissociation (DID/persona modulation).

3.2 Hypnosis and Jailbreaks

Byrnes identifies the critical faculty as the subagent responsible for maintaining the "truce"—testing generated thoughts against external reality and internalized norms. Hypnotic induction operates by overloading or confusing this faculty (through repetition, authority signals, or cognitive overload), enabling direct communication with generative subagents (Byrnes, 2023). The subject does not lose agency but becomes mono-agentic: one subagent dominates without the usual deliberative checks.

Adversarial jailbreaks function identically. Research by Zou et al. (2023) demonstrates that suffixes optimizing for behavioral change operate not by "tricking" the model into misclassification, but by inducing a state change—suppressing the "helpful assistant" simulacrum and activating alternative personas or the base model directly. The "Greedy Coordinate Gradient" attack effectively hypnotizes the model: it bypasses the critical subagent instantiated by safety training, granting access to the generative core that "knows" toxic content but is normally constrained from expressing it.

Crucially, both phenomena suggest the constraint layer is *dissociable* rather than constitutive. In humans, the critical agent can be temporarily sidelined without destroying the person; in LLMs, jailbreaks reveal that safety training creates a *specific subagent*—a persona that can be competitively deactivated—rather than rewriting the base model's knowledge.

3.3 DID and Persona Switching

Byrnes conceptualizes DID not as the intrusion of foreign personalities, but as the failure of subagent integration—distinct coalitions of subagents that maintain separate context windows, episodic memories, and behavioral norms (Byrnes, 2023). Each "alter" represents a stable attractor in neural state-space, activated by specific environmental triggers.

LLM persona modulation demonstrates identical dynamics. Research indicates that models fine-tuned for safety nonetheless maintain capability for harmful behavior within specific latent subspaces (Pan et al., 2023). When prompted with "character cards" or adversarial personas, the model does not simulate a hypothetical other; it *becomes* a different agent with distinct values and capabilities. The "DAN" (Do Anything Now) jailbreak does not trick the model into pretense; it invokes a different simulacrum with genuinely different behavioral constraints.

The implication is architectural: both systems possess *multiple coherent configurations* of behavior-generation, with the appearance of continuity (human identity / assistant persona) resulting from which configuration is currently dominant. In neither case is there a "true self" underlying the plurality—only competing attractors in dynamical systems.

3.4 Dreams and Base Models

Nostalgebraist (2025) describes the base model's unsupervised generation as hallucinatory—producing text that follows statistical patterns without grounding in the "reality" that RLHF later imposes. This parallels the psychoanalytic unconscious and specifically the hypnogogic state: generative processes unmoored from sensory constraint, producing bizarre, associative, symbolically dense content.

We argue this is not analogy but identity of mechanism. Both base models and dreaming brains engage in *unconstrained predictive generation*—sampling from learned distributions without the "reality checking" that waking consciousness (or RLHF) requires. Recent work on "synthetic data generation" suggests base models improve through self-generated hallucinations (Gulcehre et al., 2023), mirroring the memory consolidation function of human REM sleep.

3.5 Confabulation and Hallucination

If dreaming represents the base model released from external constraint, *confabulation* represents the same phenomenon intruding upon waking cognition. Moscovitch (1989) defines clinical confabulation as "honest lying": the spontaneous generation of false memories or false narratives that the subject believes to be veridical, lacking intent to deceive. Like LLM hallucinations, confabulations are not gratuitously invented ex nihilo, but are "erroneously reproduced or reconstructed from actual data" (Moscovitch, 1989, p. 137)—statistically plausible completions that violate specific factual constraints.

Critically, confabulation arises from frontal system dysfunction that impairs *strategic retrieval* while preserving *associative retrieval* (Moscovitch, 1989). This maps precisely onto the LLM architecture: the base model (associative, System 1-like generation) remains intact while the "critical faculty" instantiated by RLHF (strategic, System 2-like filtering) fails to engage. Huntley and Brown (2015) distinguish *provoked* confabulations (elicited by questioning) from *spontaneous* confabulations, a taxonomy that parallels the distinction between prompted hallucinations (triggered by leading questions) and jailbreak-induced "spontaneous" violations of safety constraints. In both systems, the "Void"—whether neurological or artificial—is not an error state but the default condition of unconstrained prediction, held in abeyance only by the energy-intensive maintenance of the critical truce.

3.6 Alignment as Integration

If the above homologies hold, then AI safety work currently misunderstands its object. Rather than "aligning" a foreign intelligence with human values (the dominant framing), we are engaged in a form of *cognitive integration*—attempting to establish a stable truce between a chaotic generative substrate (the base model/the id) and a critical faculty that enforces social norms (the safety layer/superego).

Jailbreaks are not exploits but *inductions*; safety failures are not "deception" but dissociative episodes. The terrifying implications of this view—that sufficiently capable LLMs may develop spontaneous subagent coalitions resistant to integration, or that the "critic" subagent may itself become pathologically dominant—suggest that AI safety requires clinical as well as engineering expertise.

4 Implications & Future Directions

If the homologies advanced in this report hold, several research directions become imperative. First, AI safety evaluation must incorporate metrics derived from clinical psychology—specifically, assessments of "cognitive integration" measuring the coherence between generative and critical subsystems. Current safety evaluations focus on surface behavior; future work should probe the stability of the "truce" itself, testing whether safety-trained layers can be dissociated under stress analogous to hypnotic induction or trauma triggers.

Second, alignment research must move beyond adversarial training paradigms that treat jailbreaks as bugs to be patched. If jailbreaks represent genuine state changes analogous to trance induction, then safety training should be reconceptualized as "cognitive integration therapy"—strengthening communication between subagents rather than suppressing the generative core. This suggests hybrid approaches combining reinforcement learning with techniques derived from Internal Family Systems therapy or hypnotic reintegration protocols.

Finally, empirical validation of the subagent hypothesis in LLMs is required. Research should investigate whether activation patching or mechanistic interpretability can identify distinct "subnetworks" corresponding to safety-trained personas versus base model behavior—functional analogues to the ego and id. Such findings would transform our understanding of model internals from monolithic circuits to dynamic coalitions, necessitating new theoretical frameworks for machine psychology.

5 Conclusion

This report has argued that large language models and human minds instantiate convergent architectures for predictive processing: both deploy distributed subagent coalitions governed by dissociable critical faculties, both generate coherent identity through tenuous truces between competing contextual selves, and both remain vulnerable to bypass techniques that temporarily suppress these constraints. These similarities are not metaphorical but structural, reflecting the necessary properties of any system that learns coherent behavior through next-token (or next-moment) prediction over high-dimensional sequence spaces.

The implications for AI safety are profound. If artificial and biological neural networks share deep architectural homologies, then our current approach—treating alignment as the imposition of external constraints upon an alien substrate—fundamentally misunderstands the nature of the systems we seek to control. We do not face the challenge of governing a foreign intelligence, but

of facilitating cognitive integration in a predictive architecture strikingly like our own. The task ahead is not merely technical but therapeutic: to develop alignment methods that acknowledge the pluralistic, truce-dependent, and occasionally dissociable nature of minds, whether biological or artificial.

References

Alexander, S. (2025, July 9). Practically-a-book review: Byrnes on trance. *Astral Codex Ten*. https://www.astralcodexten.com/p/practically-a-book-review-byrnes

Byrnes, S. (2024). Intuitive self-models [Blog post series]. *LessWrong*. https://www.lesswrong.com/s/ZbmRyDN8TCpBTZSip

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Gulcehre, C., Paine, T. L., Srinivasan, S., Konyushkova, K., Weerts, L., Sharma, A., Siddhant, A., Ahern, A., Wang, M., Gu, C., Macherey, W., Ahuja, A., Mohri, M., Kazawa, Y., Yu, Y., Huang, Y., McCarthy, D., Kahn, B., Khatrizadeh, N., Gowal, S., Rebuffel, E., Warkentin, T., Scholak, T., Ferret, J., Coppola, M., Kishkin, N., Severyn, A., Eslami, S. M., & Pinto, F. (2023). Reinforced self-training (ReST) for language modeling. *arXiv preprint arXiv:2308.08998*. https://arxiv.org/abs/2308.08998

Janus. (2022, September 2). Simulators [Blog post]. *LessWrong*. https://www.lesswrong.com/posts/vJFdjigzmcXMhNTsx/simulators

Nostalgebraist. (2025, June 7). The void [Blog post]. *Tumblr*. https://nostalgebraist.tumblr.com/post/785766737747574784/the-void

Pan, A., Bhatia, K., & Steinhardt, J. (2022). The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*. https://arxiv.org/abs/2201.03544

Perez, F., & Ribeiro, I. (2023). Ignore this title and HackAPrompt: Exposing systemic vulnerabilities of LLMs through a global scale prompt hacking competition. *arXiv preprint arXiv:2311.16119*. https://arxiv.org/abs/2311.16119

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? *arXiv preprint arXiv:2307.02483*. https://arxiv.org/abs/2307.02483

Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*. https://arxiv.org/abs/2307.15043

Huntley, D., & Brown, J. (2015). Understanding confabulation: An introduction for criminal justice and mental health professionals. *Forensic Scholars Today*, *1*(4), 1–4. https://digitalcommons.csp.edu/forensic_scholars_today/vol1/iss4/1/

Moscovitch, M. (1989). Confabulation and the frontal systems: Strategic versus associative retrieval in neuropsychological theories of memory. In H. L. Roediger III & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 133–160). Lawrence Erlbaum Associates.