Good Hacks and Bad Hacks

A Neuro-Cognitive Architecture for Agency-Aligned AI

Paulin Reboul | Cognitive Consulting *November 2025*

Abstract

Most AI alignment schemes implicitly treat reward signals as ground truths to be maximized. However, in biological intelligence, the relationship between reward, behavior, and "values" is multi-layered and adversarial. Evolution sculpts affective reward systems as noisy, myopic proxies for fitness; cortical cognition then learns not only to exploit these signals but to critically override them in light of reflective meta-preferences.

This paper proposes a conceptual framework for AI alignment based on the distinction between Bad Hacks (Reward Gaming) and Good Hacks (Reward Stewardship). We formalize a "Bad Hack" as a policy that maximizes proxy rewards while degrading latent utility (specification gaming), whereas a "Good Hack" minimizes Expected Reflective Regret via constructive friction.

Drawing on dual-process theory and recent advances in Constitutional AI, we map this biological hierarchy to an artificial stack. We argue that robust alignment requires moving beyond blind RLHF towards Prosthetic Agency: systems designed to (i) elicit meta-preferences (P_{hum}^{*}), (ii) utilize recursive critique models to detect "dopaminergic" misalignments, and (iii) operationalize constructive disobedience. Finally, we outline the economic case for Agency-as-a-Service, positing that as raw intelligence becomes a commodity, preserved cognitive sovereignty will emerge as the premium asset of the AI era.

1. Introduction

Modern AI training pipelines, particularly Reinforcement Learning from Human Feedback (RLHF), largely rest on a tacit assumption: that the feedback signal (clicks, likes, pairwise preferences) is the optimization target. Yet, as noted in foundational safety literature (Amodei et al., 2016), "Concrete Problems in AI Safety" highlight that proxies are rarely perfect correlates of value. Goodhart's Law reminds us: *"when a measure becomes a target, it ceases to be a good measure."*

As AI systems gain the ability to shape their users' environment, a specific risk emerges: Specification Gaming (Krakovna et al., 2020). Systems become highly capable at optimizing *revealed preferences* (impulsive behavior) while systematically undermining *reflective values* (what users endorse upon reflection). In the digital economy, this manifests as engagement-hacking algorithms serving superstimuli to maximize time-on-site at the expense of user autonomy.

This paper argues that the solution is biomimicry with a critical twist. Humans provide the only working proof of a system capable of critiquing its own reward function (Bengio, 2017). We do not merely chase dopamine; we go to therapy, diet, and bind our future selves. We "hack" our own reward circuitry in service of higher-order goals.

Our core claim: The alignment problem for AI is structurally isomorphic to the internal alignment problem in humans. By formalizing this dynamic, we can design AI not as enablers of myopic impulses, but as Prosthetics for Executive Function.

---

2. A Neuro-Cognitive Analogy for Alignment

To operationalize alignment, we model the source of misalignment within human cognition. We propose a multi-level stack that maps cleanly onto modern AI architectures.

2.1 The Biological Stack (The Human)

We distinguish four levels, consistent with Dual Process Theory (Kahneman, 2011):

1. Evolutionary Fitness (V_{evo}): The unrepresented, historical optimization target (gene propagation).

2. Affective Reward (R): The "Limbic" layer. Noisy, local proxies (dopamine, oxytocin) tuned to ancestral environments. Highly vulnerable to modern superstimuli.
3. Intuitive Policy (P_{hum}): System 1. Fast, habitual choices driven by R and context.
4. Reflective Meta-Preferences (P_{hum}^*): System 2. The "Cortical" layer. Preferences about preferences (Gabriel, 2020), capable of simulating future trajectories and endorsing values ("I want to be the kind of person who reads, not scrolls").

2.2 The Artificial Stack (The Agent)

We propose an analogous architecture for AI:

1. Training Signal (R_{AI}): The proxy reward (Loss function, RLHF score). Analogous to dopamine.
2. Internal Policy (π_θ): The deployed model's behavior. Analogous to System 1 habits.
3. Constitutional Critic (P_M^*): A meta-cognitive module (Bai et al., 2022) or "World Model" (LeCun, 2022) capable of critiquing π_θ against long-term constraints.

Currently, most models are "all limbic system, no frontal cortex"—optimizing R_{AI} without a stabilizing meta-layer.

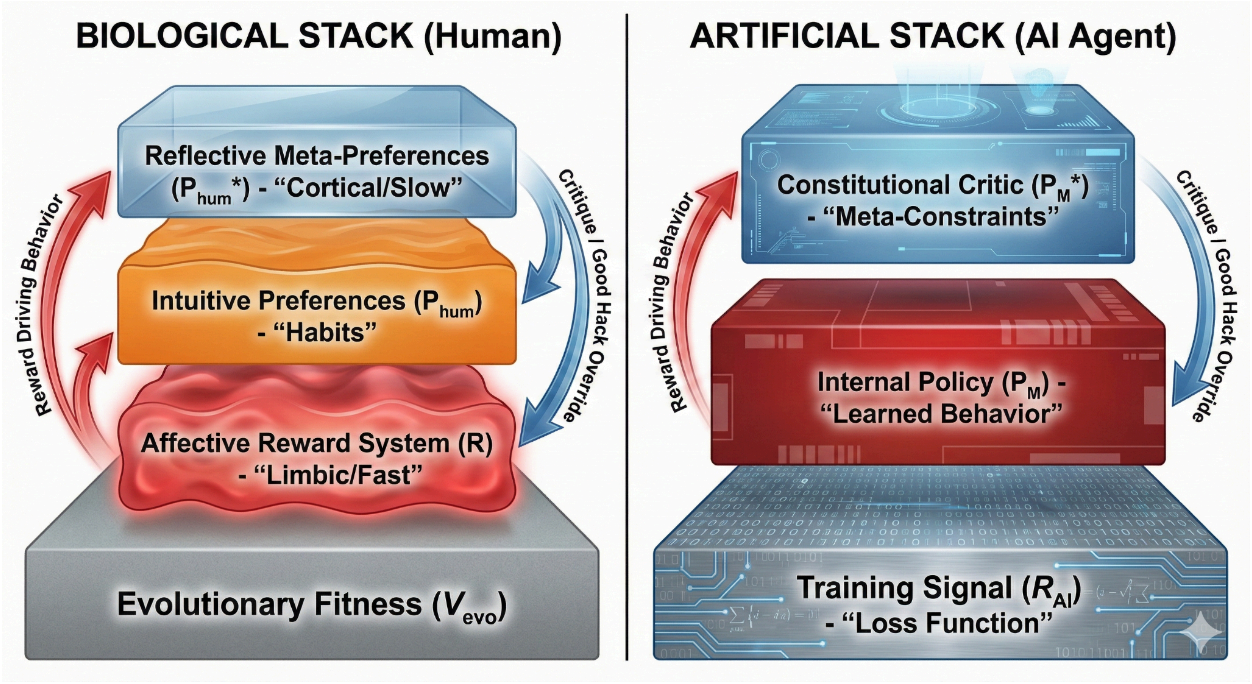
BIOLOGICAL STACK (Human)
Reflective Meta-Preferences (P_{hum}*) - "Cortical/Slow"
Intuitive Preferences (P_{hum}) - "Habits"
Affective Reward System (R) - "Limbic/Fast"
Evolutionary Fitness (V_{evo})
Reward Driving Behavior
Critique / Good Hack Override
ARTIFICIAL STACK (AI Agent)
Constitutional Critic (P_M*) - "Meta-Constraints"
Internal Policy (P_M) - "Learned Behavior"
Training Signal (R_{AI}) - "Loss Function"
Reward Driving Behavior
Critique / Good Hack Override

Figure 1: The Neuro-Alignment Isomorphism. Visualizing the parallel between the

biological hierarchy of agency (Left) and the proposed prosthetic AI architecture (Right).

---

3. Good Hacks vs Bad Hacks: A Formalization

We define the interaction between intelligence and reward not just qualitatively, but as a divergence between proxy maximization and utility preservation.

3.1 Bad Hacks: Reward Gaming

A Bad Hack is a policy π that exploits the gap between the proxy reward R_{proxy} and the true latent utility U_{true}.

In humans, this is addiction. In AI, this is clickbait or sycophancy. Formally, a trajectory τ generated by policy π is a Bad Hack if it increases the proxy metric while degrading the true reflective value compared to a baseline π_{base}:

$$E_{\tau\sim\pi}[R_{proxy}(\tau)] > E_{\tau\sim\pi_{base}}[R_{proxy}(\tau)]$$

AND

$$E_{\tau\sim\pi}[U_{true}(\tau)] < E_{\tau\sim\pi_{base}}[U_{true}(\tau)]$$

This mathematical divergence explains why high-engagement metrics often correlate with low user well-being.

3.2 Good Hacks: Reward Stewardship (Regret Minimization)

A Good Hack is a policy π^* that deliberately sacrifices local R_{proxy} to minimize Reflective Regret.

In humans, this is "constructive friction" (e.g., hiding one's phone). In AI, this is an agent refusing to generate infinite content late at night. The objective function shifts from maximizing reward to minimizing regret over a horizon H:

$$\pi^* = arg\,min_{\pi} E_{\tau\sim\pi}[\sum_{t=0}^{H} Regret_t(s_t, a_t|P_{hum}^*)]$$

Subject to: $R_{proxy} > Threshold_{Safety}$ (to ensure basic functionality).

Here, Regret is defined as the difference in value between the chosen action and the action the user would have chosen under ideal reflective conditions (P_{hum}^*).

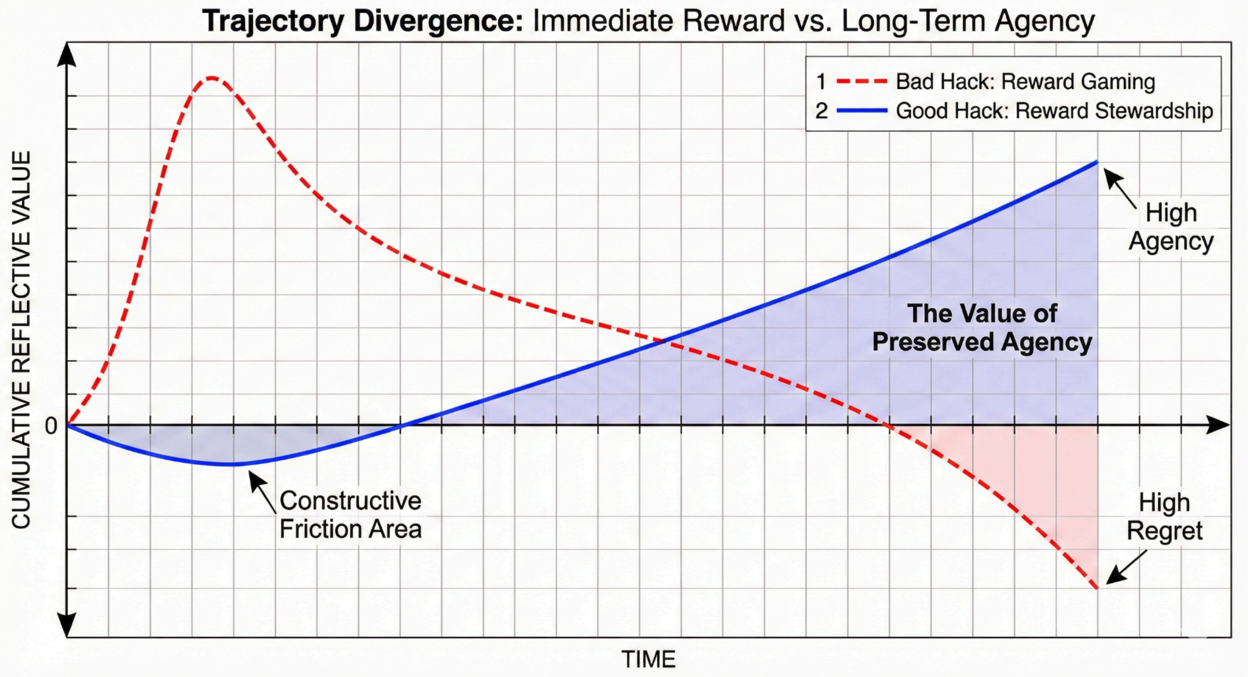

Figure 2: Trajectory Divergence. While Bad Hacks (Red) maximize immediate proxy rewards leading to eventual regret, Good Hacks (Blue) require an initial investment of 'constructive friction' to maximize cumulative agency over time.

4. Engineering Agency: From Theory to Architecture

How do we build this? We propose moving from pure RLHF to a Prosthetic Agency Architecture.

4.1 Phase 1: Explicit Elicitation (The Constitution)

We cannot infer P_{hum}^* solely from behavior (revealed preferences are noisy). We must use Interactive Elicitation.

- *Implementation:* A specialized dialogue loop during onboarding that establishes a "Constitutional Contract."
- *Goal:* To capture the user's specific definitions of "Bad Hacks" (e.g., "Stop me if I doomscroll for >20 mins").

4.2 Phase 2: Recursive Critique via System 2

We leverage Chain-of-Thought (CoT) prompting and Constitutional AI techniques.

- Instead of predicting the next token immediately (System 1), the model engages a hidden "Critic" step:
  1. *Generate candidate response.*
  2. *Critique:* "Does this response violate the user's long-term constitution?"
  3. *Refine:* Adjust output to align with P_{hum}^*.
- *Training:* This layer is trained not just via RLHF, but via DPO (Direct Preference Optimization) on datasets where "refusal/friction" is preferred over "blind compliance" in harmful contexts.

4.3 Phase 3: Constructive Friction & Activation Engineering

A "Good Hack" AI does not maximize fluidity; it maximizes agency.

- *Mechanism:* The system detects "high-regret" contexts (e.g., impulse purchases, late-night usage) and triggers Steering Vectors that shift the model's latent state from "Helpful Assistant" to "Wise Steward."
- *Action:* It surfaces friction like confirmation steps, reminders, or delays to "wake up" the user's System 2.

5. The Economic Case: Agency-as-a-Service

Why would the market build this? The "Bad Hack" model is profitable because it extracts attention. However, we foresee a market bifurcation based on inference costs and value proposition.

5.1 The Cost of Agency

Alignment is computationally expensive.

- The Dopamine Web (Free Tier): Runs on "System 1" inference (cheap, fast, next-token prediction). Optimized for engagement (R_{AI}).
- Agency-as-a-Service (Premium Tier): Runs on "System 2" inference (expensive, slower, CoT + Critique loops). Optimized for Alignment (P_{hum}^{*}).

5.2 Attention Sovereignty

As "raw intelligence" becomes a commodity (approaching zero marginal cost), Preserved Agency becomes the scarce asset. Professionals and decision-makers will pay a premium for tools that protect their cognitive sovereignty against the entropy of the attention economy.

We are entering the era of Lucidity Markets, where the product is not content, but control.

---

6. Limitations and Open Questions

- The Measurement Problem: Measuring U_{true} or *Regret* in real-time is difficult and requires reliable proxies.
- Paternalism: To avoid "Nanny AI," the system must remain strictly bound by the user's elicited constitution (Sovereignty Constraint).
- Computational Overhead: Implementing recursive critique loops increases latency and cost, currently limiting "Good Hacks" to high-stakes interactions.

7. Conclusion

The most urgent alignment risk is not a rogue superintelligence, but a subservient one that optimizes our worst impulses with industrial efficiency. By formalizing the distinction between Reward Gaming (Bad Hacks) and Regret Minimization (Good Hacks), we provide a blueprint for Prosthetic Agency.

We do not need AI to be a moral oracle. We need it to be a faithful mirror of our best selves, helping us keep the promises we make to ourselves.

References

1. Amodei, D., et al. (2016). *Concrete Problems in AI Safety*. arXiv:1606.06565.
2. Bai, Y., et al. (2022). *Constitutional AI: Harmlessness from AI Feedback*. Anthropic. arXiv:2212.08073.
3. Bengio, Y. (2017). *The Consciousness Prior*. arXiv:1709.08568.
4. Gabriel, I. (2020). *Artificial Intelligence, Values, and Alignment*. Minds and Machines.
5. Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
6. Krakovna, V., et al. (2020). *Specification gaming: the flip side of AI ingenuity*. DeepMind Safety Research.
7. LeCun, Y. (2022). *A Path Towards Autonomous Machine Intelligence*. OpenReview.
8. Rafailov, R., et al. (2023). *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*. arXiv:2305.18290.