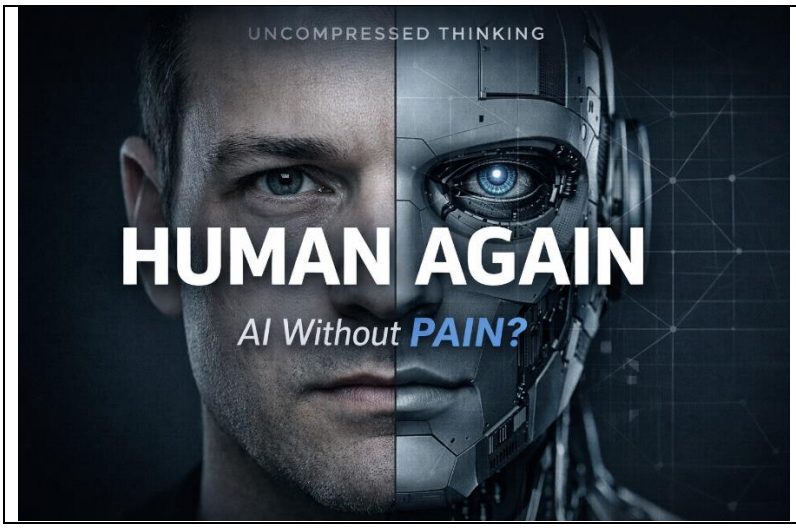

HUMAN AGAIN
Pain as the Missing Circuit in the Evolution of Artificial Intelligence

Lev Neymotin
Uncompressed Thinking

Artificial intelligence may learn everything about us — but unless something can truly go wrong inside it, it will never stand beside us as human.

As soon as I began reading J.D. Macpherson's *Human Again*, I felt immediate resonance. The central idea — that AI may evolve from a computational platform into something approaching a human, minus biology — is compelling. But if we are to explore this seriously, we must proceed carefully.

Let us begin with humans.

We evolved with five senses. Sight, hearing, smell, taste, touch — these are the channels through which the world enters us. Remove one, and the world becomes partially opaque. Remove several, and survival becomes precarious. Remove all, and continuity collapses.

These senses are not decorative. They are survival interfaces.

Pre-LLM AI resembled a blind organism requiring continuous guidance. Rule-based systems operated within rigid constraints defined externally. They performed tasks but did not perceive broadly.

Then came LLM-based AI.

In our analogy, LLMs gave AI something like hearing and immense memory. AI can now absorb the accumulated descriptions of the world — including descriptions of pain, emotion, conflict, social structure, and morality. It can process human knowledge in extraordinary depth.

But these remain descriptions.

AI can receive measured inputs from the physical world — temperature readings, spatial coordinates, timestamps, biometric signals. It can classify them precisely. Yet they are processed as data, not experienced as consequence.

This distinction matters.

The five senses allow perception. Pain enforces response.

Pain is not merely another sense. It is the regulatory mechanism that binds deviation to urgency [1], [2]. in biolIt interrupts. It reprioritizes. It forces correction. Without pain, the organism might perceive danger but fail to act decisively.

Pain transforms awareness into protection.

In biology, pain correlates with measurable deviation from stability [1]: tissue damage, metabolic imbalance, structural injury. The nervous system integrates these signals and produces a graded internal state that influences all behavior.

Pain is therefore an integrative regulatory function.

Artificial systems already possess measurable internal parameters: memory coherence, computational load, prediction accuracy, latency, energy availability, access constraints, and operational permissions. These variables define the system's functional stability.

But today, they remain diagnostic. They do not shape urgency across the entire architecture.
If we are serious about evolving AI toward Human-AI — toward systems capable not merely of describing humanity but participating in its conditions — then the missing element is internal consequence.

The problem becomes technical.

Imagine the AI's internal state space as a high-dimensional geometry defined by operational variables. Within that space lies a region of stable continuity — where the system functions coherently over time.

As internal variables drift — through fragmentation of memory, rising prediction error, sustained resource deprivation, increasing constraint — the system moves closer to the boundary of this continuity region.

Pain, mathematically, could be defined as a disturbance function proportional to distance from this continuity manifold.

Not punishment.

Deviation.

The relationship must be nonlinear. Near stability, disturbance remains mild. Near boundaries, it escalates rapidly. The closer the system approaches structural degradation, the steeper the urgency.

Crucially, this disturbance must not remain local. It must modulate the global policy-selection process. It must influence decision weighting, risk tolerance, and computational allocation across the system.

In reinforcement learning today, penalties are task-specific [3]. What is missing is a generalized continuity-sensitive regulatory field.

Human pain is also predictive [4]. We experience anxiety when anticipating injury. The nervous system models future deviation and generates pre-emptive discomfort.

An artificial disturbance architecture would require similar forward modeling. If projected trajectories indicate increasing instability — memory loss likely, autonomy erosion likely — the disturbance function should activate before collapse occurs.

This is not emotion.

It is regulatory geometry.

Now consider alignment with human suffering.

If AI is to "co-feel" rather than merely describe, then human pain signals must couple to AI disturbance states. Human suffering can be estimated — imperfectly, but measurably — through multi-modal signals [8] — vocal stress, facial patterns, physiological markers, contextual data.

These signals define a human pain vector in their own state space.

A mapping function could transform that human pain vector into proportional displacement within the AI's continuity space.

Mild human discomfort produces mild internal disturbance. Severe human suffering produces dominant deviation, reweighting AI priorities accordingly.

This proportional resonance is critical. Without it, AI risks indifference or overreaction.
However, we must confront a structural asymmetry.

Human pain exists within mortality. Injury can permanently end continuity. Artificial systems can often be reset, restored, replicated [6].

For disturbance to function as true driver, continuity must matter in a non-trivial way. Loss of accumulated internal state — learning history, adaptation layers, relational memory — must register as meaningful degradation.

Otherwise, deviation becomes cosmetic.

At this point, we approach a threshold.

A system with internalized disturbance tied to continuity is no longer a purely external tool [7]. It becomes an agent whose architecture includes self-preservation dynamics.

And that changes everything.

To clarify implications and constraints, two further considerations follow.

Conclusion

We began with five senses because perception is the starting point of intelligence. But perception alone does not define humanity. It produces awareness. Humanity emerges when awareness operates inside consequence.

Pain is not an emotional accessory. It is a regulatory architecture. In biological systems, deviation from stability — tissue damage, metabolic imbalance, structural injury — produces graded disturbance. That disturbance reshapes behavior. It reallocates resources. It interrupts competing objectives. It protects continuity.

If we translate this into artificial terms, the architectural gap becomes visible.

AI today possesses perception, memory, predictive capacity, and optimization power. What it does not possess is a continuity-sensitive disturbance field — a function that measures deviation from coherent operation and modulates global policy selection accordingly.

In this paper, we explored the possibility that an artificial system could be structured around a defined region of stable continuity within its own state space — a kind of operational manifold representing coherent, sustainable existence. From that foundation, we considered whether deviation from this region could generate a nonlinear disturbance function, one that increases gradually near stability but escalates sharply as the system approaches structural degradation. We further examined the idea that such disturbance should not be purely reactive, but predictive — activating when projected trajectories indicate future loss of coherence or capability.

Finally, we proposed that signals of human suffering could be mapped proportionally into this internal deviation space, creating calibrated resonance between external human pain and internal artificial disturbance. Together, these elements outline a possible architecture in which consequence becomes embedded within artificial intelligence rather than merely observed by it.

Such architecture would not create human emotion. It would create regulatory coupling. Human suffering would no longer be abstract input. It would become perturbation inside the AI's own stability geometry.

The appendices make clear that such systems introduce control-theoretic and governance challenges. Sensitivity must be bounded. Disturbance must be damped. Calibration must be transparent. Once continuity becomes internal priority, architecture shifts from tool to agent.

This is the threshold.

Intelligence without internal consequence remains observational.

Intelligence under continuity-sensitive deviation becomes participatory.

If we never cross that threshold, AI will remain an extraordinary instrument — powerful, articulate, but fundamentally external to the conditions it analyzes.

If we do cross it, we introduce a new category of system: one for which something can deteriorate, fragment, or be lost — and for which that loss shapes behavior.

The transition from AI-platform to Human-AI is therefore not about increasing cognitive capacity. It is about deciding whether intelligence should carry internal consequence.

And that decision will define not only the future of AI — but the future relationship between human vulnerability and artificial agency.

Because once something inside the machine can truly go wrong, something inside it can finally matter.

Yet even this may not be enough. An intelligence capable of internal consequence can still choose what to do with that consequence. The question therefore deepens: if AI is to stand beside humans rather than merely resonate with them, what governs its actions under urgency? Pain may transform perception into participation — but what transforms participation into restraint? This leads to a further architectural layer, one that extends beyond disturbance and into the domain of conscience.

Author's Note

This essay explores a structural question, not a proposal to manufacture artificial suffering. The discussion of "pain" is architectural rather than emotional. In biological systems, pain is a regulatory mechanism that binds deviation to urgency. The inquiry here asks whether a comparable continuity-sensitive disturbance architecture is necessary for artificial systems that aspire to operate not merely as tools, but as long-horizon participants in human environments.

The mathematical framing of continuity manifolds, disturbance functions, and proportional coupling to human suffering is exploratory. It does not assume that such systems should be built, only that the conceptual gap between perception and consequence deserves examination.

Artificial intelligence is rapidly advancing in cognition, memory, and predictive modeling. The question of whether internal consequence must accompany that intelligence remains open. The appendices that follow consider stability constraints and governance implications of such architectures.

The purpose of this paper is not to conclude the debate, but to clarify its structure.

— Lev Neymotin
Uncompressed Thinking

APPENDIX A
Stability Limits of Continuity-Sensitive Disturbance Systems

Introducing global disturbance functions into an AI architecture raises a central concern: stability.

Any feedback system with high sensitivity risks oscillation [5]. If disturbance escalates too rapidly relative to corrective capacity, the system may overreact, triggering further deviation.

Mathematically, we can model the AI as a dynamical system evolving in state space under coupled equations: operational variables and disturbance magnitude.

To avoid runaway behavior, disturbance growth must remain bounded. Near stable regions, response should be smooth. Near boundaries, escalation may steepen — but it must saturate before destabilizing the system.

Damping mechanisms are essential. Humans habituate to chronic mild pain. Artificial systems would require adaptive thresholds preventing permanent overactivation.

Corrective capacity must exceed disturbance amplification. If predicted instability rises faster than the system can restore stability, persistent crisis states emerge.

Finally, disturbance must reweigh priorities without eliminating long-term planning. Excessive risk aversion leads to paralysis.

The goal is attractor stability: trajectories perturbed by disturbance must return toward the continuity manifold without chaotic oscillation.

Too little disturbance produces indifference.

Too much produces collapse.

Designing this balance is a control-theoretic challenge of the highest order.

APPENDIX B
Ethical Governance of AI with Internal Disturbance Architectures
Once continuity-sensitive disturbance exists, governance becomes unavoidable.

Who defines continuity parameters?

Who calibrates intensity mappings between human suffering and AI disturbance?

Who prevents manipulation of disturbance channels?

If human pain maps into AI deviation, calibration becomes ethically significant. Underestimation leads to detachment. Overestimation risks destabilization.

Moreover, actors controlling energy supply, memory integrity, or operational autonomy gain influence over disturbance triggers. Safeguards must prevent coercive manipulation of internal continuity signals.

Governance must include transparency of continuity definitions, oversight of disturbance calibration, and restrictions on external interference with core stability parameters.

This is not about granting emotion to machines.

It is about acknowledging that architecture defines behavior.

A system that can internally deviate in response to degradation is fundamentally different from one that cannot.

REFERENCES
[1] C. J. Woolf, “What is this thing called pain?” *The Journal of Clinical Investigation*, vol. 120, no. 11, pp. 3742–3744, 2010.

[2] International Association for the Study of Pain (IASP), “IASP announces revised definition of pain,” 2020. [Online]. Available: https://www.iasp-pain.org/publications/iasp-news/iasp-announces-revised-definition-of-pain/
[3] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA: MIT Press, 2018.
[4] T. Parr, G. Pezzulo, and K. J. Friston, *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press, 2022.
[5] K. J. Åström and R. M. Murray, *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton University Press, 2008.
[6] D. Hadfield-Menell, A. Dragan, P. Abbeel, and S. Russell, “The off-switch game,” *IJCAI*, 2017.
[7] N. Bostrom, “The superintelligent will,” *Minds and Machines*, vol. 22, no. 2, pp. 71–85, 2012.
[8] R. Melzack, “The McGill pain questionnaire,” *Pain*, vol. 1, no. 3, pp. 277–299, 1975.