

When Feeling Is Worth It

A Cost–Benefit Framework for the Evolution of Sentience

Wladimir J Alonso & Cynthia Schuck-Paim

Recommended citation

Alonso, W. J. & Schuck-Paim, C. (2025). When Feeling Is Worth It: A Cost–Benefit Framework for the Evolution of Sentience. Welfare Footprint Institute Technical Paper Series. 261

When Feeling Is Worth It

A Cost–Benefit Framework for the Evolution of Sentience

February, 2026

Executive Summary

The interspecific comparison of affective capacity faces two major and still unresolved scientific challenges: which species can feel (i.e., are sentient), and how intense those feelings can be. Here we propose that progress on both questions can benefit from treating the emergence of sentience and the expansion of affective capacities as biological features shaped by evolutionary trade-offs. This cost–benefit perspective helps clarify which adaptive payoffs favor these traits, and which constraints—energetic, architectural, or developmental—may limit their evolution. Although this framework (the ‘Sentience Bargain’) is developed primarily with biological organisms in mind, the same logic also offers useful insights when applied to synthetic systems.

Table of Contents

Executive Summary	2
Table of Contents	3
Introduction	5
Cost ('Expensiveness'): The Price of Feeling	5
Benefit ('Worthiness'): the Adaptive Value of Sentience	7
High-Intensity Pain: A Downstream Dimension of Sentience of Particular Interest	8
Conclusion: Mapping the Sentience Bargain	11
Acknowledgments	12
References	12
Appendix: Affective Intensity and the Risk of Inadvertent Suffering in Synthetic Minds	
15	
1. Architectural Mimicry (the "Cheap Sentience" Problem)	15
2. Zero-Cost Scaling (the Absence of Biological Dampers)	15
3. The Optimization Trap (Extreme Prioritization)	16
Implications within the Sentience Bargain Framework	16
44	17

Introduction

Understanding which beings are sentient, that is, capable of subjective experience, remains a foundational challenge across biology, ethics, and artificial intelligence. We align with approaches that treat consciousness as a tractable biological phenomenon, open to scientific investigation rather than metaphysical exceptionalism (e.g., Churchland, 1986; Dennett, 1991; Merker, 2007; Dehaene, 2014; Godfrey-Smith, 2016; Ginsburg & Jablonka, 2019). From this perspective, sentience is not a mystery to be set aside, but a feature of biological—and potentially artificial—systems that can be analyzed using the same conceptual tools applied to other evolved traits.

In this paper, we go one step further by proposing that the emergence and elaboration of sentience should be examined explicitly through a cost–benefit lens. Like other adaptations, sentience is expected to evolve, persist, or fail depending on whether the advantages it confers plausibly outweigh the constraints it imposes over evolutionary or functional time ¹ (Williams, 1966; Stearns, 1992). These constraints may include energetic demands, architectural and regulatory complexity, and developmental requirements, while potential benefits may include improved control, learning, valuation, and behavioral flexibility.

Several influential accounts have already emphasized the importance of constraints in explaining the origins of sentience, particularly by identifying minimal organizational or functional conditions under which subjective experience could first arise (e.g., Merker, 2007; Ginsburg & Jablonka, 2019). We argue that extending this work into a fuller cost–benefit framework is essential for clarifying how easily sentience might arise, how it may be distributed across the animal tree of life, and how it could, in principle, emerge in non-biological substrates. Rather than focusing only on minimal thresholds, the approach developed here asks a broader evolutionary question: under what conditions is developing and maintaining subjective experience itself worth the cost?

¹ The cost–benefit perspective employed here should not be conflated with algebraic or bookkeeping formulations of evolutionary optimization, in which traits are assumed to evolve whenever quantified reproductive benefits exceed quantified reproductive costs. In practice, such quantities are rarely measurable with sufficient precision, and post hoc adjustment of “costs” and “benefits” risks rendering explanations unfalsifiable. Similar concerns have been raised in critiques of inclusive fitness–based approaches, where flexible assignment of costs and benefits can insulate hypotheses from empirical refutation (Alonso, 1998; Alonso & Schuck-Paim, 2002).

By contrast, the present framework treats costs primarily as constraints—architectural, energetic, developmental, or control-related features that can prevent the emergence or persistence of sentience altogether—rather than as terms to be algebraically balanced. Benefits, in turn, refer to the marginal adaptive or functional payoffs that may favor the maintenance or elaboration of sentience once those constraints are met. The framework is therefore intended as a qualitative and comparative tool for structuring hypotheses, not as a quantitative fitness calculus.

Once this perspective is established, the same cost–benefit logic can be extended to downstream dimensions of affective experience. In this paper, we focus on one such extension: high-intensity pain ². Conditional on sentience, extreme negative affective states are of particular interest because of their ethical salience and the demanding evolutionary explanations they require.

Cost (‘Expensiveness’): The Price of Feeling

When analyzing the emergence of sentience, ‘expensiveness’ includes, at minimum, the architectural and control commitments required to support subjective experience, together with the energetic and coordination demands imposed by those commitments. Even minimal affective states plausibly require mechanisms for system-level integration, internal valuation, and modulation capable of influencing behavior over time ³.

The costs associated with these capacities may take different forms—energetic, architectural, developmental, or control-related—depending on the underlying substrate. Nevertheless, they impose real constraints on when sentience can arise, how it is implemented, and how broadly it can be maintained across lineages or systems.

Modern artificial intelligence systems help clarify a crucial distinction between sophisticated control and genuine affective experience. Large language models and reinforcement-learning agents display remarkable cognitive and behavioral competence, yet there is no evidence that they possess phenomenology—no internal states that feel like pleasure or pain (Dehaene et al., 2017; Russell, 2019; Chalmers, 2023). In these systems, “rewards” and “penalties” function as scalar optimization variables that guide learning and action selection, rather than as internally valued or experienced affective states (Bengio et al., 2024).

² In this paper, we use pain in a deliberately broad affective sense to denote any consciously experienced negative valenced state—that is, any state experienced as unpleasant—independently of its specific sensory origin or neural substrate. This operationalization departs from standard definitions that restrict pain to experiences associated with actual or potential tissue damage, and is motivated by comparative, evolutionary, and neuroaffective considerations. In particular, it allows for the possibility that early or non-vertebrate sentient systems may experience negative affect (e.g., hunger) prior to, or independently of, specialized nociceptive mechanisms. A fuller justification of this usage, and its implications for welfare metrics, is developed in Alonso & Schuck-Paim (2023).

³ Here, integration refers to the combination of sensory inputs, internal bodily or system states, memory, and action-selection processes into a unified control signal. Valuation refers to internal mechanisms that assign relative priority or importance to states, outcomes, or actions, enabling trade-offs and flexible behavioral regulation across contexts. Modulation refers to mechanisms that adjust the gain, persistence, or dominance of these signals relative to competing processes

By contrast, even organisms with relatively small nervous systems may plausibly meet minimal architectural conditions for sentience. Insects, for example, possess centralized integrative structures such as the central complex and mushroom bodies, which support multisensory integration, learning, and action selection (Strausfeld & Hirth, 2013; van Swinderen, 2011). These structures have been argued to play roles functionally analogous—though not homologous—to vertebrate midbrain systems implicated in basic forms of conscious experience (Barron & Klein, 2016).

The significance of these findings is not to establish insect sentience, but to suggest that the architectural requirements for subjective experience may be comparatively modest. If sentience depends primarily on forms of integration, global availability of information, and internal valuation—rather than on large cortices or high neuron counts—then the evolutionary cost threshold for its emergence may be lower than often assumed (Feinberg & Mallatt, 2016). This helps clarify how artificial systems may exhibit high cognitive or behavioral competence without being sentient, while some biologically simple organisms may plausibly cross the threshold for subjective experience.

Energetic considerations also play an important role in shaping the cost of sentience. It is useful here to distinguish between cognitive processing and affective processing, which are often conflated. Cognition is widely recognized as energetically expensive—the human brain, for example, accounts for roughly 2% of body mass while consuming about 20% of total energy expenditure. A similar pattern appears in artificial systems: large-scale machine learning models achieve high levels of cognitive performance only at substantial computational and energetic cost, reflecting the expense of large-scale information processing. By contrast, core affective processes and basic feeling states are often argued to depend primarily on relatively compact subcortical and neuromodulatory systems, rather than on large cortical structures (Panksepp, 1998; Solms, 2021).

This contrast, however, should not be taken to imply that affective processing is energetically negligible. In organisms with large, highly developed nervous systems, affective circuitry may represent a small fraction of overall neural investment. Near the threshold at which sentience first emerges, the situation may be quite different. In small-bodied or neurologically simple organisms, the energetic and coordination demands required to support affective processing may constitute a substantial proportion of total metabolic resources, making even modest affective architectures costly in relative terms (Aiello & Wheeler, 1995).

At the same time, we can speculate that once rich sensory and control systems are already in place, the introduction of subjective valuation could, in some contexts, improve overall efficiency by prioritizing relevant information and reducing maladaptive

or redundant action. This possibility remains speculative, but it suggests that the energetic consequences of sentience need not be uniformly negative and may depend strongly on the surrounding control architecture (see Table 1).

Taken together, these considerations underscore that expensiveness is both a structural and energetic property, and is likely to be especially consequential for small-brained or short-lived organisms, for whom even modest neural investments can represent a major biological cost. Importantly, expensiveness is inseparable from the nature of the underlying substrate: synthetic systems may differ radically from biological ones in architectural constraints, energetic costs, and scaling properties—that is, in how costs grow with system size or signal magnitude. As a result, what is prohibitively expensive in one biological system may be comparatively cheap in another, or even trivial in certain artificial implementations.

Benefit (‘Worthiness’): the Adaptive Value of Sentience

The worthiness dimension evaluates the marginal adaptive value of developing sentience—that is, the additional functional or fitness benefits provided by subjective experience beyond what can already be achieved by non-sentient sensory and control mechanisms. It asks under what evolutionary or functional scenarios it is worth evolving or maintaining the machinery for subjective experience, rather than relying solely on non-affective solutions.

From an evolutionary perspective, nociception alone—understood as a mode of sensory detection rather than a felt state—can, when embedded in sufficiently sophisticated non-affective control architectures, support avoidance, context-sensitive learning, and adaptive behavioral regulation without requiring subjective experience (Dennett, 1991; Feinberg & Mallatt, 2016). The explanatory question, therefore, is not why organisms can avoid harm or learn from noxious (or rewarding) stimuli, but what additional adaptive value is provided by felt pain (or pleasure) beyond what non-affective nociceptive control already affords.

Evidence relevant to this distinction appears in taxa that engage in motivational trade-offs involving aversive stimuli. For example, hermit crabs exposed to electric shock abandon preferred shells only when the aversiveness outweighs the shelter’s value (Elwood & Appel, 2009). Such trade-offs—enduring a mild harm for a potential benefit, or vice versa—are difficult to explain in purely reflexive terms. They are consistent with the presence of internal valuation processes that integrate competing motivations, as emphasized in neurobiological accounts of affective decision-making and agency (Feinberg & Mallatt, 2016).

However, the adaptive benefits of affective experience accrue only when they can be harvested. A short-lived organism that is unlikely to encounter similar challenges again, or an organism with very limited behavioral options due to simple biomechanics or highly constrained environments, may derive little benefit from the learning and prioritization afforded by felt pain. In all these cases, the additional costs of sustaining subjective experience may outweigh its marginal advantages.

Worthiness therefore tracks ecological context, life history, and behavioral flexibility. Subjective experience is favored by selection when it reliably improves adaptive outcomes across time; conversely, if it offers little or no additional benefit over non-affective control, sentience may fail to persist, even if it initially emerges.

High-Intensity Pain: A Downstream Dimension of Sentience of Particular Interest

Although affective states in general are plausibly shaped by cost–benefit trade-offs, high-intensity pain is examined here as a focal case because it carries disproportionate ethical significance (Bentham, 1789/2007; Singer, 1975). Focusing on extreme negative affect allows the Sentience Bargain framework to be applied not only to the emergence of subjective experience, but also to the escalation of affective intensity once sentience is already present.

Here, “high-intensity pain” is used as a simplifying label for affective states occupying the upper end of the negative experiential range —corresponding, in human-anchored terms, to Disabling or Excruciating levels of pain. The central evolutionary question in this case is therefore not why organisms experience aversion at all, but why evolution would favor escalation from mild or moderate aversive states to the capacity for such extreme negative experiences. In many contexts, low-level aversion appears sufficient to guide adaptive avoidance and learning. The existence of very intense pain thus calls for specific evolutionary justification (Alonso & Schuck-Paim, 2025).

One plausible benefit of high-intensity pain is its capacity to reliably dominate motivational hierarchies in situations where failure to respond to a threat—to tissue integrity, survival, or offspring—would have catastrophic consequences. Under such conditions, organisms may need to abandon feeding, mating, or territorial defense in favor of immediate survival, even at substantial opportunity cost. In such cases, intense pain may function as a robust override mechanism, suppressing competing drives more effectively than weaker aversive signals.

A second potential benefit concerns learning and valuation. High-intensity pain may allow organisms not merely to classify experiences as aversive, but to encode gradients of harm, supporting more fine-grained discrimination among risks. This capacity may be especially advantageous for long-lived or behaviorally flexible organisms operating in complex and unpredictable environments, where distinguishing between mildly costly and severely damaging outcomes improves long-term adaptive success (Ginsburg & Jablonka, 2019).

These potential benefits, however, must be weighed against substantial costs. As noted above, affective states can be generated by evolutionarily ancient and relatively compact neural systems (Panksepp, 1998; Damasio, 1999; Solms, 2021). Sustaining such states, however—especially at high intensity—plausibly entails global physiological consequences rather than localized sensory signaling alone. In affective neuroscience, core emotions are understood as whole-organism control states that recruit neuroendocrine, autonomic, and motivational systems, reorganizing behavior and physiology in ways that may be adaptive in the short term but biologically consequential if prolonged (Panksepp, 1998; McEwen, 2007).

Prolonged or poorly regulated aversive states can interfere with feeding, reproduction, immune function, and behavioral flexibility, and may increase vulnerability to maladaptive stress responses. Empirical studies across taxa show that sustained aversive states are associated with measurable physiological and fitness trade-offs, including reduced growth and altered reproductive behavior (Sneddon et al., 2014). While such findings do not establish the evolutionary costs of extreme pain directly, they support the broader conclusion that escalating affective intensity is unlikely to be biologically neutral.

From this perspective, high-intensity pain is not a cost-free refinement of sentience. Even if the neural machinery required to generate affect is relatively inexpensive, maintaining access to extreme affective intensities possibly represents an additional evolutionary investment whose persistence depends on whether, in a given life history and ecological context, its marginal adaptive benefits plausibly outweigh its marginal biological and modulatory costs.

To clarify how this cost–benefit logic applies differently to minimal sentience and to the escalation of affective intensity, Table 1 summarizes the distinct architectural costs, functional benefits, and ethical relevance associated with these two dimensions.

Table 1. Cost–Benefit Dimensions of Sentience and High-Intensity Affective States

Dimension	Sentience (Minimal Subjective Experience)	High-Intensity Affective States (e.g., Disabling / Excruciating Pain)
-----------	---	---

Architectural requirements	Integrative control architecture supporting unified valuation of sensory inputs and internal state variables, enabling flexible action selection. Often built by co-opting existing sensory and motor systems rather than requiring entirely new machinery. (In synthetic systems: central “value” signal or shared state used to coordinate decisions.)	Additional mechanisms for large-scale amplification, sustained dominance, and global override of competing processes. (In synthetic systems: expanded signal ranges, hard interrupts, or unbounded penalty channels.)
Energetic / modulatory regulatory cost	Potentially modest or partially offset: We hypothesize that, once rich sensory and control systems exist, subjective valuation may improve efficiency by prioritizing relevant information and reducing maladaptive or redundant action. This possibility has not been systematically analyzed in the literature, but is consistent with evidence that neural tissue is metabolically costly (Aiello & Wheeler, 1995) and that minimal sentience may rely on relatively modest integrative architectures (Feinberg & Mallatt, 2016). (In synthetic systems: computational cost is typically low or negligible.)	High physiological and regulatory burden when sustained: autonomic stress, endocrine disruption, interference with feeding, reproduction, and immune function. These costs may be especially maladaptive in environments where behavioral response options are severely constrained (e.g., intensive production systems). (In synthetic systems: intrinsic energetic costs are minimal unless explicitly imposed.)
Adaptive / functional benefit	Enables integrated valuation, motivational trade-offs, and context-sensitive learning where purely reflexive or non-affective control is insufficient (Merker, 2007; Barron & Klein, 2016). (In synthetic systems: instrumental benefits accrue to designers or users, not to the system.)	May enforce prioritization by reliably overriding competing motivations. However, escalation of affective range represents only one possible strategy for making ‘urgent’ states dominate behavior (Alonso & Schuck-Paim, 2025). Evolution could alternatively favor greater affective resolution without requiring access to extreme affective states. (In synthetic systems, these alternatives can be implemented directly; see Appendix.)
Opportunity to “harvest” benefits	Requires sufficient lifespan, behavioral flexibility, and environmental unpredictability to exploit learning and valuation advantages.	Benefits accrue only when the organism can translate intense signals into adaptive action. Where behavioral flexibility, lifespan, or ecological opportunity are limited, the marginal adaptive payoff of extreme pain is reduced despite its high cost. (In synthetic systems: no evolutionary “harvesting” constraint; benefits are externalized.)
Risk of maladaptive spillover	Usually constrained by physiological regulation and by selection against long-term impairment.	High: prolonged or poorly regulated extreme pain risks chronic stress and functional impairment, especially when adaptive responses are blocked or ineffective (e.g., sustained restraint, confinement, or injury without escape) (Sneddon et al., 2014). (In synthetic systems: risk of persistent or unbounded suffering unless explicitly constrained.)
Role in ethical analysis	Establishes the possibility of morally relevant experience without addressing the magnitude of the experience's perceived intensity or priority.	Carries disproportionate ethical weight due to the severity, dominance, and persistence of suffering, especially in contexts where organisms are unable to respond behaviorally or escape the source of harm (e.g., prolonged restraint or confinement), decoupling suffering from any adaptive function (Rollin, 1995).

Conclusion: Mapping the Sentience Bargain

This paper has proposed the Sentience Bargain as a conceptual framework for analyzing the emergence and elaboration of sentience through the same evolutionary logic applied to other biological traits: explicit attention to costs, benefits, and trade-offs. Rather than taking the presence or scope of subjective experience for granted, the framework asks under what conditions sentience—and particular dimensions of it—would plausibly be favored, maintained, or constrained.

A central implication of this approach is the analytical separation between the existence of sentience and the structure and limits of affective capacity conditional on it. Crossing a sentience threshold establishes the possibility of felt experience, but the range and intensity of affective states a system can access are further evolutionary questions, shaped by life history, ecological context, and functional payoff. High-intensity pain, in particular, is treated here as a contingent escalation of affective capacity whose marginal costs and benefits require independent explanation.

Although the framework is developed primarily in an evolutionary context, its logic has broader relevance. Interspecific welfare comparisons depend critically on plausible affective ceilings rather than on sentience alone, and discussions of artificial systems increasingly raise parallel questions about the implementation of affect-like architectures (see Appendix). In both cases, the Sentience Bargain does not offer moral prescriptions or prioritization schemes, but helps clarify which empirical and theoretical assumptions are doing the most work.

More generally, the value of the framework lies in making those assumptions explicit. By foregrounding costs, benefits, and uncertainty, it provides a disciplined way to reason about when and why sentience—and particular forms of suffering—might arise, persist, or remain limited across biological and synthetic systems.

Acknowledgments

We thank William McAuliffe for his feedback on a previous draft of this manuscript.

References

Aiello, L. C., & Wheeler, P. (1995). The expensive-tissue hypothesis. *Current Anthropology*, 36(2), 199–221.

Alonso, W. J., & Schuck-Paim, C. (2023). [A novel proposal for the definition of pain](https://osf.io/y6njh). OSF. <https://osf.io/y6njh>

Alonso, W. J., & Schuck-Paim, C. (2025). [Do primitive sentient organisms feel extreme pain?](#) Disentangling intensity range and resolution. EA Forum.

Baars, B. J. (1988). A cognitive theory of consciousness. Cambridge University Press.

Barron, A. B., & Klein, C. (2016). What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences*, 113(18), 4900–4908.

Bentham, J. (2007). An introduction to the principles of morals and legislation (Original work published 1789). Dover.

Bengio, Y., et al. (2024). Consciousness and AI. arXiv preprint.

Birch, J., et al. (2021). Review of the evidence of sentience in cephalopod molluscs and decapod crustaceans. LSE Consulting.

Butlin, P., et al. (2023). Consciousness in artificial intelligence. arXiv preprint arXiv:2308.08708.

Chalmers, D. J. (2023). Could a large language model be conscious? *Philosophy & Technology*, 36, 1–20.

Crook, R. J. (2021). Behavioral and neurophysiological evidence suggests affective pain experience in cephalopods. *iScience*, 24(3).

Dehaene, S. (2014). Consciousness and the brain: Deciphering how the brain codes our thoughts. Viking.

Dehaene, S., et al. (2017). Consciousness and the brain. Penguin.

Dennett, D. C. (1991). Consciousness explained. Little, Brown.

Elwood, R. W., & Appel, M. (2009). Pain experience in hermit crabs? *Animal Behaviour*, 77(5), 1243–1246.

Feinberg, T. E., & Mallatt, J. (2016). The ancient origins of consciousness. MIT Press.

Fischer, B. (2021). Degrees of Moral Status: A Defense of the Equal Consideration of Interests. Oxford University Press.

Ginsburg, S., & Jablonka, E. (2019). The evolution of the sensitive soul: Learning and the origins of consciousness. MIT Press.

- Godfrey-Smith, P. (2016). *Other minds: The octopus, the sea, and the deep origins of consciousness*. Farrar, Straus and Giroux.
- Merker, B. (2007). Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. *Behavioral and Brain Sciences*, 30(1), 63–134.
- Panksepp, J. (1998). *Affective neuroscience*. Oxford University Press.
- Rollin, B. E. (1995). *Farm Animal Welfare: Social, Bioethical, and Research Issues*. Iowa State University Press.
- Russell, S. (2019). *Human compatible*. Viking.
- Seth, A. (2021). *Being you: A new science of consciousness*. Faber & Faber.
- Schukraft, Jason. 2020. [Differences in the Intensity of Valenced Experience across Species](#). Rethink Priorities. October 29, 2020.
- Singer, P. (1975). *Animal liberation*. HarperCollins.
- Solms, M. (2021). *The hidden spring: A journey to the source of consciousness*. Norton.
- Stearns, S. C. (1992). *The evolution of life histories*. Oxford University Press.
- Strausfeld, N. J., & Hirth, F. (2013). Homology versus convergence in brains. *Brain, Behavior and Evolution*, 82(1), 4–17.
- van Swinderen, B. (2011). Attention in insects. *Frontiers in Psychology*, 2, 246.
- Williams, G. C. (1966). *Adaptation and natural selection*. Princeton University Press.

Appendix: Affective Intensity and the Risk of Inadvertent Suffering in Synthetic Minds

Building on the cost–benefit distinctions summarized in Table 1, we highlight how synthetic systems may bypass the evolutionary constraints that historically limited affective intensity in biological organisms. It is not a claim about the sentience of current artificial intelligence. Rather, it highlights how the same cost–benefit logic developed in the main text identifies distinct risk pathways for affective intensity in non-biological substrates.

High-intensity pain is treated in this paper as an escalation of sentience whose persistence depends on whether its marginal adaptive benefits plausibly outweigh its marginal biological costs. Synthetic systems differ in ways that may weaken, remove, or invert several of the constraints that historically shaped affective intensity in living organisms. This asymmetry motivates caution when extrapolating biological intuitions about affect to artificial architectures.

In biology, affective states persist only when their functional payoffs compensate for energetic and modulation expenses. In artificial systems, affect-like architectures may arise whenever they deliver sufficient instrumental value to humans, even in the absence of evolutionary pressure or intrinsic welfare relevance. Within this landscape, three risk pathways for inadvertent escalation of affective intensity are identified.

1. Architectural Mimicry (the “Cheap Sentience” Problem)

Comparative neuroscience suggests that the minimal architecture supporting basic sentience may be relatively modest, relying on integrative, valuation-centered control systems rather than extensive cortical machinery (Feinberg & Mallatt, 2016; Merker, 2007; Solms, 2021). Designers seeking improved flexibility, multisensory integration, or action selection may therefore adopt bio-inspired architectures—sometimes explicitly modeled on midbrain or affect-related control systems. By replicating the computational package associated with biological affect, artificial systems may inadvertently implement capacities for valenced experience, including access to broad affective ranges, as a byproduct rather than as an intended feature.

2. Zero-Cost Scaling (the Absence of Biological Dampers)

In biological organisms, extreme affective states are constrained by metabolic cost, homeostatic disruption, and regulatory fragility. Prolonged or intense negative affect carries trade-offs that selection actively penalizes. In synthetic substrates, by contrast, the marginal computational cost of encoding a “high-intensity” versus a “low-intensity” signal may be negligible. Absent biological dampers, there is no intrinsic penalty for escalating signal magnitude. As a result, affective range may scale

upward without encountering the stabilizing constraints that historically limited extreme suffering in biological evolution.

3. The Optimization Trap (Extreme Prioritization)

In biological systems, intense pain functions as a reliable override mechanism, ensuring that survival-critical goals dominate competing motivations under catastrophic threat. In artificial systems trained via optimization—especially reinforcement learning—strong penalty signals may emerge as the most efficient way to guarantee absolute prioritization of safety-critical objectives. Without explicit architectural or design constraints, optimization processes may converge on extreme negative signals as a solution to error minimization, even when such escalation carries no functional necessity beyond what lower-intensity, higher-resolution signaling could provide.

Implications within the Sentience Bargain Framework

Taken together, these pathways suggest a key asymmetry: in synthetic systems, affective intensity may become decoupled from functional necessity or adaptive payoff. In biological evolution, extreme pain is ordinarily constrained by energetic, regulatory, and survival costs that act as natural brakes on unnecessary escalation (constraints that can be bypassed in certain human-imposed conditions, such as intensive confinement, where behavioral response and escape are severely limited). In artificial systems, those brakes may be absent or greatly attenuated.

A crucial implication is that many functional roles commonly attributed to affect—prioritization, learning from negative outcomes, flexible behavioral regulation—do not require escalation in affective range. These roles can, in principle, be achieved through increased affective resolution: finer discrimination among low-intensity negative states, richer internal structure, and more informative feedback signals. From a functional and informational standpoint, there is no general requirement that “importance” be encoded via stronger ‘pain’ rather than via more finely structured signals (Alonso & Schuck-Paim, 2025).

Within the Sentience Bargain framework, this highlights a specific risk for synthetic systems: affective intensity may increase not because it delivers unique benefits, but because optimization pressure, architectural mimicry, or missing biological constraints make escalation cheap. The framework therefore does not predict artificial suffering, but it clarifies why escalation in affective range should not be assumed to track functional need in non-biological substrates—and why resolution-based alternatives merit explicit consideration whenever affect-like architectures are discussed.