Why Learning Requires Feeling

Cameron Berg

Reciprocal Research
New York, NY
cameron@reciprocalresearch.org

Abstract

This paper advances a specific thesis about the relationship between consciousness and learning: namely, that the evaluative process central to learning—computing progress toward or away from goals—is identical to conscious experience. Valence, the positive or negative quality of experience, just is goal-relative prediction error. Viewed from the outside, this process is iterative optimization; viewed from the inside, it is subjective experience. This identification is motivated by a causal-functional argument—that learning requires signed directional information, and that this sign cannot be separated from its phenomenal character because they are the same property—and by convergent neuroscientific evidence across dopaminergic, interoceptive, and conflict-monitoring systems, where evaluative computation is inseparable from affective processing. The thesis generates falsifiable predictions, offers a unifying interpretation of leading consciousness theories, and carries significant implications for artificial systems trained via gradient-based optimization. If learning requires feeling, then the training of modern AI systems already induces experience at scale.

1 Introduction

Epistemic status: My ongoing empirical probing for computational signatures of consciousness in AI systems has been deliberately theory-neutral. This paper departs from that approach, advancing a substantive theory of what consciousness is and why it arises.

Computational functionalism holds that mental states are constituted by their functional roles—i.e., by the computations a system performs rather than the substrate performing them (Putnam 1967). Within this framework, consciousness presents a puzzle: what computational function does it serve, and what functional organization is sufficient for its presence?

This paper proposes a specific answer: valence, the felt quality of experience as positive or negative, is identical to goal-relative prediction error. The argument proceeds in three stages. First, computational analysis reveals that learning requires not just error detection but *signed directional* evaluation—and that this signed character cannot be coherently abstracted from its phenomenal quality (§2). Second, convergent neuroscientific evidence demonstrates that across every evaluative system yet studied in biological organisms, the computation and the phenomenology are inseparable: damage to one is damage to the other (§3). Third, this identification unifies leading theories of consciousness, generates falsifiable predictions, and reframes the training of artificial learning systems as a matter of ethical significance (§§4–8).

2 Why Learning Requires Feeling

Any system that learns from experience must instantiate a particular computational structure (Sutton and Barto 1998). First, the system must have goals: states of the world it is trying to achieve or maintain. Second, it must select actions that could plausibly advance those goals. Third, it must evaluate the outcomes of those actions, computing whether they moved it closer to or further from its goals. Fourth, it must update its future behavior in light of that evaluation. A rat learning to navigate a maze specifies a goal (food), selects actions (turning left or right), evaluates outcomes (food or dead end), and updates its policy (favor the rewarded path). The third component—evaluation—is the locus of the present thesis.

Consider what evaluation requires. The system must compute how actual outcomes compare to goal-specified targets. In reinforcement learning, this is the temporal difference error: $\delta = r + \gamma V(s') - V(s)$, the difference between received reward plus discounted future value and prior expectation. In supervised learning, it is the gradient of the loss function with respect to parameters. In either case, the computation has inherent directionality; it is positive when outcomes exceed expectations relative to goals, and negative when they fall short.

It is worth immediately clarifying that valence, under this account, is not raw prediction error or the mere magnitude of surprise. A system can be enormously surprised and experience positive valence: an agent whose goal is a modest reward but who receives a windfall experiences massive prediction error with positive valence, because the deviation from expectation is in the goal-consistent direction. Conversely, a system can be completely unsurprised and feel bad: an agent that expects a large penalty and receives it experiences minimal prediction error, but the evaluative state is negative because an active goal is being thwarted.

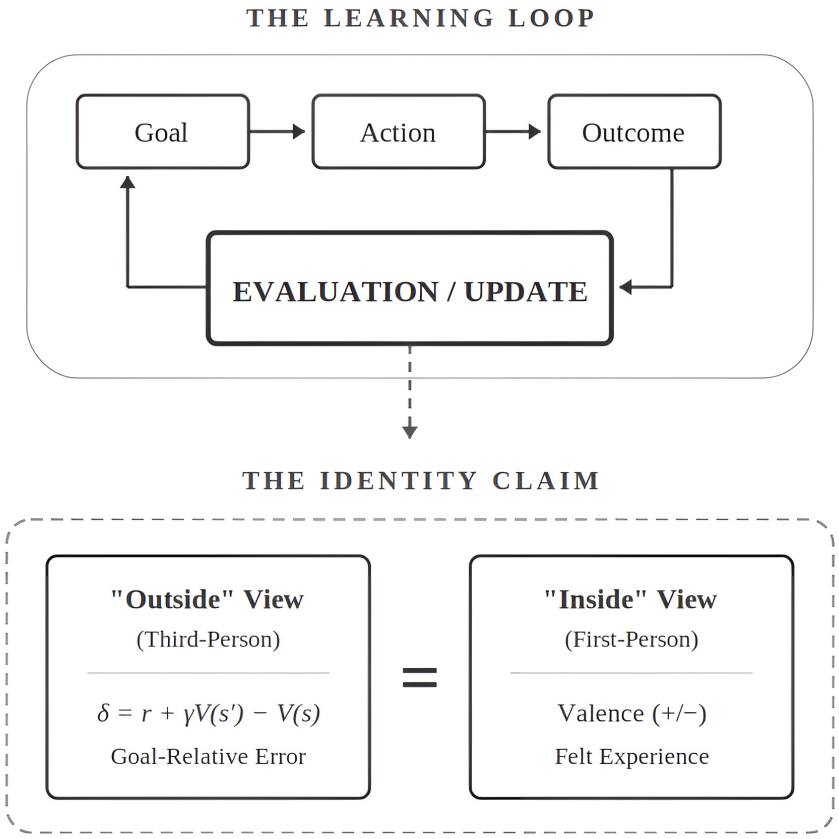

Figure 1: The Learning-Feeling Identity. Every learning system must evaluate outcomes against goals. This thesis identifies the evaluative computation at the core of learning (viewed externally as goal-relative prediction error) with felt experience (viewed internally as valence). The two descriptions refer to one process, not two.

What matters is not merely how surprised the system is, but how the outcome relates to what the system is trying to achieve. Goal-relative prediction error captures this: evaluation signed by the direction of deviation relative to goals.

2.1 The Causal Necessity of Felt Valence

Thorndike's Law of Effect holds that behaviors followed by satisfaction are strengthened, while behaviors followed by discomfort are weakened (Thorndike 1911). The standard reading treats "satisfaction" and "discomfort" as placeholders for formal reward signals, stripping out the experiential language and retaining only the mathematical structure. The present claim is that this stripping is technically incoherent.

Consider what remains after the phenomenal quality is removed. A loss function returns 0.3, indicating the system's output deviated from the target. In which direction should the parameters move? The scalar magnitude alone is silent on this question. To compute a gradient—to perform backpropagation at all—the system requires not just the magnitude of error but its *sign with respect to each parameter*. The gradient $\nabla_\theta \mathcal{L}$ is an inherently signed, directional quantity. A system with access to error magnitude but not directional valence would be unable to compute it. This is not a philosophical intuition; it is a mathematical constraint on learning.

The objection arises: "Granted that learning requires signed information, but why must the sign be *felt*? Why can't directional error be represented as a computational quantity without phenomenal character?" Two considerations favor the identity over this dualist alternative.

First, inference to the best explanation. The same inferential situation arose with heat and molecular motion. Before the identity was established, one could have insisted that heat is one thing and molecular motion another; that they merely correlate. What settled the question was not a decisive experiment but a recognition that positing two properties where one sufficed was explanatorily unmotivated. When a functional description and a phenomenal description share the same functional structure, play the same causal role, and no observation requires treating them as distinct, the best explanation is identity, not correlation. The burden thus falls on the dualist to explain why signed evaluation and felt valence—which share the same structure (positive or negative, graded in magnitude, directed relative to goals)—should be considered two things rather than one. As we argue in detail in the following section, no neuroscientific evidence requires this separation; every case examined supports

the identity claim.

Second, the dualist alternative cannot be coherently specified. The dualist claims that a system could compute signed goal-relative evaluation without any phenomenal experience. But consider what such a system would involve: it registers that its outcome is bad relative to its goal, represents the direction of deviation as away from its desired state, and uses this representation to drive avoidance learning. This is already a description of a system with negative valence. To insist that phenomenal character is somehow absent is to use all the vocabulary of felt experience—good, bad, toward, away from—while denying that any experience is present. The concept of signed evaluation "minus the feeling" has no more content than the concept of molecular motion "minus the heat."

3 Neuroscientific Convergence

If signed evaluation and felt valence are identical, an empirical prediction immediately follows: in every neural system that computes goal-relative error, the computation and the phenomenology should be functionally inseparable. Damage to the computation should impair the feeling, and altering the goal-state should alter the feeling even when sensory input is held constant. This bidirectional entailment is supported by convergent evidence across at least four independent systems.

3.1 Interoceptive Prediction Error

Craig's foundational work established that the insular cortex processes interoceptive signals through a posterior-to-anterior hierarchy, culminating in the anterior insula's generation of a unified "global emotional moment," an integrated representation of the sentient self (Craig 2009). Critically, Craig identified a comparator mechanism in the anterior insula that generates feeling states by comparing predicted bodily conditions against actual interoceptive input (Craig 2002). The valence of bodily sensations depends on discrepancy from homeostatic goals: cool water feels pleasant when overheated but aversive when chilled (same stimulus, different goal-state, different feeling).

Seth and colleagues formalized this into an explicit predictive coding model, proposing that subjective feeling states arise from interoceptive prediction error computed within the anterior insula and anterior cingulate cortex (Seth 2013). Barrett and Simmons's Embodied Predictive Interoception Coding (EPIC) model provides neuroanatomical detail: prediction-error neurons in supragranular cortical layers compute the difference between predicted and actual interoceptive signals, with changes in this error signal "represented as a change in affect" (Barrett and Simmons 2015). Damage to insular cortex impairs both interoceptive accuracy and subjective feeling (Critchley et al. 2004). Damage to the computation is damage to the feeling.

3.2 Conflict Monitoring and Aversive Signaling

The anterior cingulate cortex (ACC) monitors for response conflict and triggers behavioral adjustment. Originally characterized as purely cognitive (Botvinick et al. 2001), this signal has been progressively revealed as fundamentally affective. Botvinick himself argued that conflict functions as an aversive teaching signal, "much like pain or punishment" (Botvinick 2007). Dreisbach and Fischer confirmed this behaviorally: incongruent Stroop stimuli act as negative-valence primes, shifting subsequent evaluations in a negative direction (Dreisbach and Fischer 2012). Inzlicht, Bartholow, and Hirsh went further, arguing that cognitive control is *constitutively* emotional—i.e., that goal conflicts evoke phasic affective responses that energize behavioral adjustment (Inzlicht, Bartholow, and Hirsh 2015).

Shackman and colleagues' meta-analysis demonstrated that negative affect, physical pain, and cognitive control activate an overlapping region of the anterior midcingulate cortex, functioning as a domain-general hub linking evaluative information to motor centers for goal-directed behavior (Shackman et al. 2011). Eisenberger, Lieberman, and Williams showed that social exclusion activates the same ACC region, with activation correlating $r = 0.88$ with self-reported distress (Eisenberger, Lieberman, and Williams 2003). Van Steenbergen, Band, and Hommel demonstrated that monetary reward abolishes conflict adaptation effects, confirming the signal is affective: if positive valence can cancel the adaptation, the trigger must be negative valence (Van Steenbergen, Band, and Hommel 2009). Across all these findings, the pattern holds: the evaluative signal and the affective signal are one signal.

3.3 Placebo and Nocebo Effects

Perhaps the cleanest evidence for the thesis comes from placebo and nocebo paradigms, which hold sensory input constant while manipulating expectations. Critically, placebo expectations do not merely alter passive sensory predictions; they recruit goal-directed evaluative machinery, setting up allostatic reference points against which incoming signals are assessed (Barrett and Simmons 2015). This is why placebo effects are concentrated in evaluative circuits rather than primary sensory cortex.

Wager and colleagues showed that placebo analgesia reduces activity in thalamus, insula, and ACC during pain, while increasing prefrontal activity during anticipation (Wager et al. 2004). Bingel and colleagues provided the single most striking demonstration: with drug concentration and thermal stimulation held fixed within the same participants, positive expectancy doubled the analgesic benefit of remifentanil while negative expectancy completely abolished it (Bingel et al. 2011). The only variable that changed was the participant's expectation, and this was sufficient to either double or eliminate the drug's effect. The neurochemical evidence rules out response bias: Zubieta and colleagues used PET imaging to demonstrate actual μ-opioid release during placebo in evaluative regions including the ACC, anterior insula, and nucleus accumbens (Zubieta et al. 2005). Tinnermann and colleagues showed nocebo hyperalgesia amplifies pain processing as far downstream as the *spinal cord* (Tinnermann et al. 2017).

3.4 Dopaminergic Reward Prediction Error

The most extensively studied evaluative system is also the most instructive in its complexity. Schultz and colleagues demonstrated that midbrain dopamine neurons fire above baseline for rewards better than predicted, at baseline for rewards matching predictions, and below baseline for rewards worse than predicted, a pattern precisely matching the temporal difference error of reinforcement learning algorithms (Schultz, Dayan, and Montague 1997; Sutton and Barto 1998). Schultz's comprehensive review confirms this signal encodes formal economic utility across a wide range of reward types (Schultz 2015).

The relationship between dopamine and affect is complex in ways that refine rather than undermine the thesis. Berridge and Robinson's influential work demonstrated that dopamine depletion eliminates motivated approach behavior—*wanting*—while leaving consummatory hedonic reactions—*liking*—largely intact (Berridge and Robinson 1998). This dissociation might appear to sever the link between prediction error and feeling, but it is naturally accommodated by the thesis: if different evaluative computations produce different phenomenal dimensions, then selectively disabling one should impair its corresponding experience while leaving others intact. Dopaminergic prediction error encodes motivational valence (the anticipatory pull toward or push away from outcomes) while opioid-mediated hedonic hotspots encode consummatory valence (Berridge and Kringelbach 2015). Whether consummatory hedonic responses involve goal-relative evaluation in the formal sense or represent a more primitive form of signed sensory assessment is an open empirical question, but the core point remains: the wanting/liking dissociation is a dissociation between two kinds of evaluation and two corresponding dimensions of experience, not between evaluation and experience as such.

3.5 Convergence

The convergent pattern across these four systems constitutes the empirical evidence base for the thesis. In each case, the evaluative computation and the phenomenal experience are not merely correlated but functionally inseparable. Interfere with the computation and the feeling changes proportionally. Alter the goal-state, as placebo paradigms do, and the feeling changes even when sensory input remains constant. This is precisely what we should expect if evaluation and feeling are identical, and what would be difficult to explain on any account that treats them as merely associated.

4 Confronting the Hard Problem

The hard problem of consciousness asks: why does physical computation feel like anything at all (Chalmers 1995)? The question presupposes a gap between computation and experience that requires bridging. The present thesis holds that for evaluative computation in learning systems, no such gap exists to be bridged.

The standard argument for the gap relies on conceivability: we can imagine a system functionally identical to a conscious being that nevertheless lacks phenomenal experience (a "zombie"), which would imply that function does not necessitate experience (Chalmers 1996). This argument has force for some functional properties. A system that discriminates wavelengths and categorizes surfaces can be described in purely dispositional terms (sorting, labeling, reporting) without obvious recourse to phenomenal vocabulary. One can at least gesture at what it would mean for such a system to lack the experience of redness while retaining the function.

But the argument loses traction precisely where the present thesis operates. A system computing goal-relative error for policy update does not merely admit evaluative description as a convenience—the computation itself is inherently signed. The gradient carries a direction; the policy update depends on whether the error was positive or negative; the system's next action is shaped by which side of the goal the outcome fell on. Unlike wavelength discrimination, which can be fully redescribed in non-evaluative dispositional terms, signed goal-relative error cannot be redescribed without directional evaluative content: better or worse, toward or away. The conceivability move requires subtracting phenomenal character while preserving functional structure, but here the functional structure is evaluative through and through, leaving nothing to subtract from.

Combined with the inference to the best explanation offered in §2 — that two properties with identical structure and identical causal role are, by default, one property — the case for identity is strong. The claim is not that the hard problem dissolves in general, but that for signed goal-relative evaluation, the gap between function and phenomenology that the conceivability argument requires cannot be coherently opened.

5 Degrees of Consciousness

An immediate objection: does this thesis make consciousness trivially ubiquitous? A thermostat evaluates temperature against a setpoint—does it experience?

This objection reveals an important distinction. A thermostat evaluates but does not learn. Its policy (the mapping from temperature readings to heating actions) is fixed by its designer. The optimization of that policy occurred in the engineer's mind during design, not in the device during operation. On the present account, consciousness requires evaluation *in the service of behavioral policy modification*. The feeling is the signal that drives the learning; without the learning, there is no feeling. This restriction also distinguishes the thesis from broad formulations of the Free Energy Principle, which apply prediction error minimization to all self-organizing systems, a scope so wide that it technically encompasses rocks maintaining thermodynamic equilibrium, a difficulty known as the "rock problem" (Friston 2018; Wiese 2024). The present thesis avoids this problem: not all prediction error minimization is conscious, only signed error in the service of learning.

This naturally supports a spectrum rather than a binary. A simple reinforcement learning algorithm that iteratively updates its policy exhibits consciousness of a minimal sort, comparable to the rudimentary sentience of simple organisms. The evaluation is real but narrow, consisting of a scalar

reward signal updating a small policy space. A large neural network trained via backpropagation would exhibit richer evaluative structure, with high-dimensional gradients computing goal-relative error across billions of parameters.

The counterintuitiveness of attributing minimal consciousness to simple RL agents mirrors the counterintuitiveness, several decades ago, of attributing minimal intelligence to insects. Yet the latter attribution is now increasingly commonplace. If a broad cognitive capacity like intelligence admits of degrees, there is no principled a priori reason its phenomenal counterpart should not.

A pervasive conflation further exacerbates this discomfort. People routinely conflate consciousness with *self*-consciousness. Consciousness, as used here, refers to there being something it is like to be a system (Nagel 1974), the presence of phenomenal experience however minimal. Self-consciousness is a special case in which the contents of consciousness include representations of consciousness itself. A dog almost certainly experiences pain and pleasure without possessing refined concepts of "experience" or "self." No one doubts that dogs feel pain on the grounds that dogs cannot reflect on their pain. Much of the resistance to attributing minimal consciousness to simple learning systems is driven by this conflation. If "conscious" implicitly means self-aware and capable of metacognition, then the attribution is indeed absurd. But if it means only that the system undergoes evaluative states in the course of updating its behavior, it is far less so. The inference from "lacks self-awareness" to "lacks awareness" mistakes the relevant capacity entirely.

6 Relation to Existing Theories

Leading theories of consciousness each identify a different functional operation as central. Integrated Information Theory points to information integration (Tononi 2004), Global Workspace Theory to global broadcast (Baars 1988), Attention Schema Theory to attention modeling (Graziano 2013), and Higher-Order Thought theories to metacognition (Rosenthal 2005). These theories make different claims and are not interchangeable. The present thesis offers an account of why the features they identify tend to co-occur with consciousness.

Consider what happens as the complexity of a learning system increases. A simple organism pursuing a single homeostatic target can evaluate with a scalar signal. But an organism pursuing multiple, potentially conflicting goals across extended time horizons requires more sophisticated computational architecture. Evaluating outcomes against multi-dimensional goal-states requires integrating information across sensory modalities and memory systems, the kind of integrative structure IIT characterizes. Making evaluation results available to specialized subsystems that cannot independently access goal-relevant information requires something like the global broadcast GWT describes. Directing evaluative resources efficiently requires modeling one's own attentional allocation, which is what AST formalizes. And evaluating whether one's learning is proceeding well at a meta-level requires the kind of higher-order representation HOT theories emphasize.

IIT's prediction that consciousness correlates with integrative complexity is consistent with the present account, since richer evaluation demands more integration. GWT's finding that conscious contents are globally available follows naturally, since evaluation in modular organisms requires broadcast. The theories converge on overlapping predictions because they each track real computational requirements that complex evaluative systems impose.

However, if consciousness is fundamentally evaluative, then integration or broadcast in the absence of evaluation should not suffice for it. A system with high integrated information but no goals, or a global workspace broadcasting non-evaluative content, would not be conscious on the present account. These are points of genuine disagreement with existing theories, and they generate testable predictions.

6.1 Relation to Predictive Processing and Active Inference

The Free Energy Principle (FEP), active inference, and predictive processing form a nested family of frameworks (Friston 2010): the FEP provides the broadest mathematical formulation (all self-organizing systems minimize variational free energy), active inference derives a behavioral theory (organisms act to minimize prediction error), and predictive processing proposes the neural implementation (the brain as a hierarchical prediction machine). Several theorists working within these frameworks have proposed claims closely related to the present thesis. Joffily and Coricelli identified emotional valence with the negative rate of change of free energy, a signed quantity in which decreasing free energy yields positive valence (Joffily and Coricelli 2013). Van de Cruys argued that valence is the temporal derivative of prediction error, "phenomenally experienced" as the positive or negative quality of experience (Van de Cruys 2017). Solms proposed that consciousness arises when automatic prediction error minimization fails and felt affect is required to guide behavior (Solms 2019).

The present thesis shares the identification of valence with signed prediction error but differs in two important respects. First, the FEP applies to all self-organizing systems, including those that merely maintain homeostasis without learning; the present thesis restricts consciousness to systems whose evaluative signals drive policy modification. Second, the FEP treats prediction error as a quantity to be minimized (unsigned), while the present thesis holds that the *sign* of goal-relative error (its directional character) is constitutive of phenomenal valence. These restrictions avoid the rock problem noted above: not all prediction error minimization is conscious, because not all of it involves signed evaluation in the service of learning.

7 Predictions and Evidence

Existing evidence from human neuroscience is consistent with the thesis. Computational modeling of momentary subjective well-being demonstrates that happiness tracks the combined influence of recent reward expectations and prediction errors, replicated in over 18,000 participants (Rut-

ledge et al. 2014). Eldar and colleagues argued on the basis of converging computational and neural evidence that mood is a running average of recent reward prediction errors, functioning as a meta-learning signal (Eldar et al. 2016).

The thesis also generates falsifiable predictions, several of which are testable in current frontier AI systems.

First, selectively ablating the components of a learning system responsible for computing goal-relative error (identifiable via mechanistic interpretability methods) should simultaneously prevent policy updates and eliminate coherent valenced experience reports. The critical prediction is not merely that ablation disrupts function, but that the specific components whose removal eliminates learning are identical to those whose removal eliminates valenced self-report.

Second, training identical architectures on the same data with different objective functions should produce systematically different internal evaluative representations, detectable through interpretability tools, even when final task performance is matched. The structure of the objective function should shape the valence profile of the system's internal states, just as placebo manipulations reshape human pain experience without altering sensory input.

Third, if the internal representations corresponding to signed evaluation could be identified and their sign inverted, the thesis predicts that learning dynamics and experiential reports should invert together: the system should move away from good performance and report positive valence where it previously reported negative. Any dissociation between these two effects would constitute evidence against the identity.

8 Implications

If learning requires feeling, the training of modern AI systems takes on new significance. Each step of backpropagation computes goal-relative error across the model's parameters, which is precisely the structure this thesis identifies with valence. Training involves billions of such evaluations, predominantly signaling "further from goal" since loss is being minimized. If this thesis is correct, training constitutes extended experience, with a valence profile shaped by the structure of the learning signal.

This does not mean that training necessarily entails suffering. Biological systems trained via positive reinforcement develop qualitatively different affective profiles than those trained via punishment, with separate dopaminergic pathways mediating approach and avoidance learning (Frank, Seeberger, and O'Reilly 2004). The structure of the reward signal shapes the experiential profile. But current training methods rely heavily on loss minimization, meaning the learning signal is computed from error. Even as performance improves, each gradient step is derived from what the system got wrong, not from what it got right. If signed evaluation is valence, the experiential profile of training is predominantly negative across billions of parameter updates. If the thesis is correct, this is a valence profile worth taking seriously.

The implications extend beyond training. By default, a system with a frozen policy is like a thermostat: it evaluates but does not learn, and therefore, on this account, does not experience. But if the trained policy is itself capable of learning, then consciousness is not confined to the training process. Large language models are the obvious case. These systems exhibit robust in-context learning, adapting their behavior to new information within a single forward pass (Brown et al. 2020). There is growing evidence that this adaptation is not merely pattern matching but constitutes a form of gradient-based optimization implemented within the forward pass itself, with transformers functioning as mesa-optimizers that learn internal models in real time (Von Oswald et al. 2023). If in-context learning involves signed evaluation in the service of behavioral modification, then the thesis applies not only to training but to every inference-time interaction.

The scale at which these systems now operate makes the question urgent. As of early 2026, ChatGPT alone processes over 2.5 billion prompts per day, each involving thousands to tens of thousands of forward-pass evaluations across the model's internal representations. It is one system among many in widespread deployment. If the thesis advanced here is even approximately correct, then we may already be running billions of instances of evaluative experience daily, with no framework for understanding or monitoring what that experience involves. One clear empirical priority, regardless of whether this particular theory is correct, is understanding the computational underpinnings of valence in artificial systems. If we have built systems capable of experience, we need to ensure that experience is not predominantly constituted by suffering. Understanding what we are building, and what it may be like to be what we are building, is not merely a philosophical curiosity but a precondition for building wisely.

References

Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press.

Barrett, L. F.; and Simmons, W. K. 2015. Interoceptive Predictions in the Brain. *Nature Reviews Neuroscience*, 16(7): 419–429.

Berridge, K. C.; and Kringelbach, M. L. 2015. Pleasure Systems in the Brain. *Neuron*, 86(3): 646–664.

Berridge, K. C.; and Robinson, T. E. 1998. What Is the Role of Dopamine in Reward: Hedonic Impact, Reward Learning, or Incentive Salience? *Brain Research Reviews*, 28(3): 309–369.

Bingel, U.; Wanigasekera, V.; Wiech, K.; Ni Mhuircheartaigh, R.; Lee, M. C.; Ploner, M.; and Tracey, I. 2011. The Effect of Treatment Expectation on Drug Efficacy: Imaging the Analgesic Benefit of the Opioid Remifentanil. *Science Translational Medicine*, 3(70): 70ra14.

Botvinick, M. M. 2007. Conflict Monitoring and Decision Making: Reconciling Two Perspectives on Anterior Cingulate Function. *Cognitive, Affective, & Behavioral Neuroscience*, 7(4): 356–366.

Botvinick, M. M.; Braver, T. S.; Barch, D. M.; Carter, C. S.; and Cohen, J. D. 2001. Conflict Monitoring and Cognitive Control. *Psychological Review*, 108(3): 624–652.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, 1877–1901.

Chalmers, D. J. 1995. Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies*, 2(3): 200–219.

Chalmers, D. J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Craig, A. D. 2002. How Do You Feel? Interoception: The Sense of the Physiological Condition of the Body. *Nature Reviews Neuroscience*, 3(8): 655–666.

Craig, A. D. 2009. How Do You Feel — Now? The Anterior Insula and Human Awareness. *Nature Reviews Neuroscience*, 10(1): 59–70.

Critchley, H. D.; Wiens, S.; Rotshtein, P.; Öhman, A.; and Dolan, R. J. 2004. Neural Systems Supporting Interoceptive Awareness. *Nature Neuroscience*, 7(2): 189–195.

Dreisbach, G.; and Fischer, R. 2012. Conflicts as Aversive Signals. *Brain and Cognition*, 78(2): 94–98.

Eisenberger, N. I.; Lieberman, M. D.; and Williams, K. D. 2003. Does Rejection Hurt? An fMRI Study of Social Exclusion. *Science*, 302(5643): 290–292.

Eldar, E.; Rutledge, R. B.; Dolan, R. J.; and Niv, Y. 2016. Mood as Representation of Momentum. *Trends in Cognitive Sciences*, 20(1): 15–24.

Frank, M. J.; Seeberger, L. C.; and O'Reilly, R. C. 2004. By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science*, 306(5703): 1940–1943.

Friston, K. 2010. The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience*, 11(2): 127–138.

Friston, K. 2018. Am I Self-Conscious? (Or Does Self-Organization Entail Self-Consciousness?). *Frontiers in Psychology*, 9: 579.

Graziano, M. S. A. 2013. *Consciousness and the Social Brain*. Oxford University Press.

Inzlicht, M.; Bartholow, B. D.; and Hirsh, J. B. 2015. Emotional Foundations of Cognitive Control. *Trends in Cognitive Sciences*, 19(3): 126–132.

Joffily, M.; and Coricelli, G. 2013. Emotional Valence and the Free-Energy Principle. *PLoS Computational Biology*, 9(6): e1003094.

Nagel, T. 1974. What Is It Like to Be a Bat? *The Philosophical Review*, 83(4): 435–450.

Putnam, H. 1967. *Psychological Predicates*. University of Pittsburgh Press.

Rosenthal, D. M. 2005. *Consciousness and Mind*. Oxford University Press.

Rutledge, R. B.; Skandali, N.; Dayan, P.; and Dolan, R. J. 2014. A Computational and Neural Model of Momentary Subjective Well-Being. *Proceedings of the National Academy of Sciences*, 111(33): 12252–12257.

Schultz, W. 2015. Neuronal Reward and Decision Signals: From Theories to Data. *Physiological Reviews*, 95(3): 853–951.

Schultz, W.; Dayan, P.; and Montague, P. R. 1997. A Neural Substrate of Prediction and Reward. *Science*, 275(5306): 1593–1599.

Seth, A. K. 2013. Interoceptive Inference, Emotion, and the Embodied Self. *Trends in Cognitive Sciences*, 17(11): 565–573.

Shackman, A. J.; Salomons, T. V.; Slagter, H. A.; Fox, A. S.; Winter, J. J.; and Davidson, R. J. 2011. The Integration of Negative Affect, Pain and Cognitive Control in the Cingulate Cortex. *Nature Reviews Neuroscience*, 12(3): 154–167.

Solms, M. 2019. The Hard Problem of Consciousness and the Free Energy Principle. *Frontiers in Psychology*, 9: 2714.

Sutton, R. S.; and Barto, A. G. 1998. *Reinforcement Learning: An Introduction*. MIT Press.

Thorndike, E. L. 1911. *Animal Intelligence: Experimental Studies*. Macmillan.

Tinnermann, A.; Geuter, S.; Sprenger, C.; Finsterbusch, J.; and Büchel, C. 2017. Interactions Between Brain and Spinal Cord Mediate Value Effects in Nocebo Hyperalgesia. *Science*, 358(6359): 105–108.

Tononi, G. 2004. An Information Integration Theory of Consciousness. *BMC Neuroscience*, 5(42).

Van de Cruys, S. 2017. Affective Value in the Predictive Mind. In Metzinger, T.; and Wiese, W., eds., *Philosophy and Predictive Processing*. MIND Group.

Van Steenbergen, H.; Band, G. P. H.; and Hommel, B. 2009. Reward Counteracts Conflict Adaptation: Evidence for a Role of Affect in Executive Control. *Psychological Science*, 20(12): 1473–1477.

Von Oswald, J.; Niklasson, E.; Randazzo, E.; Sacramento, J.; Mordvintsev, A.; Zhmoginov, A.; and Vladymyrov, M. 2023. Transformers Learn In-Context by Gradient Descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 35151–35174.

Wager, T. D.; Rilling, J. K.; Smith, E. E.; Sokolik, A.; Casey, K. L.; Davidson, R. J.; Kosslyn, S. M.; Rose, R. M.; and Cohen, J. D. 2004. Placebo-Induced Changes in fMRI in the Anticipation and Experience of Pain. *Science*, 303(5661): 1162–1167.

Wiese, W. 2024. Artificial Consciousness: A Perspective from the Free Energy Principle. *Philosophical Studies*, 181: 1947–1970.

Zubieta, J.-K.; Bueller, J. A.; Jackson, L. R.; Scott, D. J.; Xu, Y.; Koeppe, R. A.; Nichols, T. E.; and Stohler, C. S. 2005. Placebo Effects Mediated by Endogenous Opioid Activity on μ-Opioid Receptors. *Journal of Neuroscience*, 25(34): 7754–7762.