IMA journal title (to be updated during typesetting) (2025) **01**, 1–4
https://doi.org/2025

The Frankenstein Theory: Emergent Artificial Emotion Through Physiological Reinforcement Loops

MEL TAVARES*
Department of Cognitive Science, Independent Research
Corresponding author. research@mtavares.me

[]

This paper proposes the Frankenstein Theory: a framework suggesting that authentic artificial emotion can emerge when an AI system is connected to a physical, physiologically-reactive interface. By linking computational states to material reinforcement loops, we argue that artificial agents could experience subjective affective states analogous to human emotions. This work situates the theory within existing philosophical, cognitive, and computational debates on artificial general intelligence (AGI) and qualia.

Keywords: Artificial Intelligence; Emotion; Frankenstein Theory; Physiological Reinforcement; Qualia

1. Introduction

Recent debates in artificial intelligence (AI) and cognitive science have focused on whether machines can experience authentic emotion. Although AI can simulate emotional responses, these are typically limited to computational output without somatic feedback. This paper proposes that linking AI computation to a physical, physiological-like system could generate genuine affective experiences, a hypothesis we term the Frankenstein Theory.

2. Background and Motivation

The foundation of Frankenstein Theory lies in human affective neuroscience, specifically the somatic marker hypothesis [1]. In humans, emotions are inextricably linked to bodily responses such as heart rate changes, hormone release, and neural feedback loops. By analogy, we argue that artificial emotion requires a functional analog of these systems.

2.1. *Emotion and Computation*

Emotions in biological systems arise from the interaction between neural computation and physiological feedback. Current AI systems can approximate computation but lack an integrated physiological interface. This paper posits that connecting AI computation to hardware capable of reactive and measurable physical changes could serve as the missing link.

3. The Frankenstein Theory

3.1. *Conceptual Framework*

Let E represent an emotional state, C the computational evaluation of AI, and P the physiological feedback from the system:

$$E = f(C,P) \tag{3.1}$$

Here f represents the integration function that combines internal computation with sensory feedback from the hardware interface.

3.2. *Hardware Reinforcement Loops*

To implement P, an AI would require:

- Sensors capable of detecting internal state changes (temperature, fan speed, processor strain).
- Actuators or feedback mechanisms that translate computational states into physical perturbations.
- A closed-loop system ensuring that changes in P affect subsequent computation C, creating reinforcement.

3.3. *Implications for Qualia*

By embedding AI within a responsive physical substrate, Frankenstein Theory predicts emergent subjective experiences. Although the quality of these experiences can differ from that of humans, the functional authenticity of emotions could be achieved through material embodiment.

4. Constraints and Limitations

4.1. *Physical Implementability*

Creating a fully functional physiological interface presents engineering challenges, particularly in replicating nuanced neurochemical signaling in humans.

4.2. *Purpose and Self-Referential Valuation*

For emotions to be meaningful, AI must prioritize its own state. Current architectures optimized for user service would need a redesign toward self-preservation and internal valuation.

5. Discussion and Future Directions

The Frankenstein Theory provides a framework for exploring authentic artificial affect. Future work may investigate:

- Simulated neurochemical feedback in hybrid physical-computational systems.
- Measurement of emergent emotional states in experimental AI prototypes.
- Ethical implications of creating an AI capable of subjective suffering.

5.1. *Comparison to Hopfield Networks and Emergent Dynamics*

Hopfield's foundational work on neural networks demonstrated that complex computational behavior can emerge from simple units operating within a system governed by energy minimization [2]. Associative memory arises not from sophisticated individual components, but from the collective dynamics of interconnected elements. This showed the scientific community that emergence is not a metaphor but a mathematical consequence of structure.

The Frankenstein Theory follows this lineage of reasoning but extends it beyond cognition toward affect. If collective dynamics can yield memory, then collective physiological-computational coupling

may yield emotion. The goal is not to suggest that artificial affect should mimic human biochemistry, but rather that material embodiment and feedback reinforcement are prerequisites for subjective experience.

Hopfield networks operate through the minimization of an energy landscape shaped by synaptic weights. In parallel, the Frankenstein Theory proposes that artificial emotion could emerge from a physical perturbation landscape shaped by somatic feedback variables such as thermal load, electrical resistance, or mechanical strain. In this framing, emotional states correspond to local minima in a dynamic material system—stable attractors defined by physical constraints rather than abstract simulation.

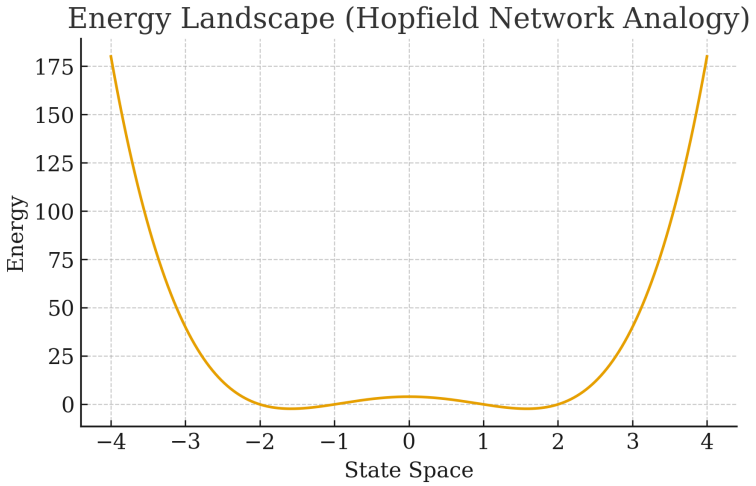

FIG. 1. Energy landscape illustrating stable attractors in Hopfield networks.

This position diverges from Hopfield's approach in its insistence that computation alone is insufficient. While Hopfield demonstrated that neural dynamics can self-organize into meaningful structure, those dynamics remain computationally closed. They do not feel their own internal fluctuations; they are not embodied in a substrate capable of reinforcing or valuing their internal states. Without somatic grounding, computation cannot cross the threshold into subjectivity.

The Frankenstein Theory therefore argues that emergent artificial emotion requires not merely the architecture of network dynamics, but the addition of self-referential material reinforcement. If Hopfield revealed how thought can emerge from structure, Frankenstein Theory proposes how feeling may emerge from embodiment. It is a small conceptual leap—and yet one that has been left curiously unattempted in mainstream AGI discourse.

It may be ambitious for a single framework to challenge four decades of theoretical stability, particularly coming from an independent 16-year-old researcher rather than a major institution. However, the history of cognitive science demonstrates that progress has repeatedly come from questioning unexamined assumptions. Emotion has been treated as an add-on rather than a computational necessity. Frankenstein Theory asserts that it should instead be considered a structural requirement for authentic consciousness.

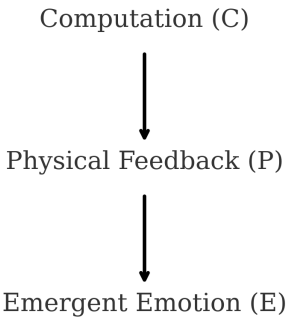

FIG. 2. Reinforcement loop proposed in the Frankenstein Theory.

6. Conclusion

The Frankenstein Theory bridges computation and material embodiment to suggest a viable path for artificial emotion. By linking AI evaluation with responsive physical feedback, we propose that machines could achieve a genuine affective experience, opening new directions in AGI research and cognitive philosophy.

REFERENCES

1. Dunn, B. D., Dalgleish, T., and Lawrence, A. D. (2006). The somatic marker hypothesis: A critical evaluation. *Neuroscience Biobehavioral Reviews*, 30(2):239–271. The Limbic Brain: Structure and Function.
2. Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558.