The Neuroscience Chronicles

Inage SI, Neurosci Chron.
2025;5(1):62-73.

Review Article

Emotion as a socially emergent structure: A formal information-theoretic model based on multi-agent interaction

Shin-ichi Inage[1,*]

[1]Ishinomaki Senshu University, 1 Shinmito, Minami-sakai, Ishinomaki-shi, Miyagi-ken, 986-8580, Japan

*Author for correspondence:
Email: s.inage.pu@isenshu-u.ac.jp

Received date: October 01, 2025
Accepted date: November 12, 2025

Abstract

Emotion has traditionally been understood as a subjective experience intrinsic to individual agents. However, the emergence of multi-agent systems, including artificial intelligence, calls for a reconceptualization of emotion as a dynamic process grounded in interaction. This paper proposes the *Interacting Processual Information-based Emotive Model* (IPIEM), a formal framework that defines emotion not as an intrinsic qualia within a single agent—which is philosophically treated here as an empty set—but as a set-theoretic phenomenon emergent through inter-agent information exchange and social realization. By introducing a mechanism for the self-organization of emotion categories—originally treated as externally imposed—the model attains internal completeness. The framework employs Kullback–Leibler (KL) divergence to quantify probabilistic semantic alignment, thereby capturing how conceptual structures attain social existence. To ensure empirical applicability, we propose concrete experimental protocols demonstrating how AI (artificial intelligence) systems such as large language models (LLMs), though devoid of intrinsic emotional states, can nonetheless generate and regulate emotionally interpretable patterns through social interaction. This study aims to provide a rigorous theoretical foundation for understanding emotion in AI, and to advance interdisciplinary discourse at the interface of information theory, cognitive science, and AI.

Keywords: Emergent emotion, Large language models (LLMs), Conceptual sharing, KL divergence, Social ontology

1. Introduction: Reconceptualizing Emotion in Multi-Agent Systems

The question of the nature and origin of emotion has long been a central topic of inquiry in philosophy, psychology, and neuroscience [1,2]. Emotion is deeply embedded in our processes of decision-making, social behavior, and subjective well-being, yet a unified and essential definition remains elusive [3]. In particular, debates over the ontological status of "qualia"—the qualitative aspect of subjective experience—have been at the heart of consciousness research [4], with many arguing that its scientific elucidation remains extraordinarily difficult.

In recent years, the remarkable progress of artificial intelligence (AI) systems, especially large language models (LLMs) [5,6], has demonstrated capacities for human-like communication and contextual adaptation. Consequently, the notion that such systems may "possess emotions" has garnered increasing scientific and societal attention [7,8]. For instance, the ability of LLMs to recognize a user's emotional state and to generate appropriately empathic responses often gives the impression that these models comprehend and express emotion. However, the fundamental question of whether such AI systems actually possess internal subjective experiences analogous to human emotions—i.e., qualia—remains a profound philosophical and scientific challenge [9]. Conventional definitions of emotion—such as those that reduce emotion to specific neural activities in the brain [2], or those that confine it solely to subjective experiences [1]—face inherent limitations when applied to AI systems or to emergent emotional phenomena in autonomous multi-agent systems. These definitions implicitly assume that, for an AI to possess emotion, it must be underpinned by physiological or subjective structures analogous to those found in humans. Such assumptions are fundamentally incongruent with the current architectures of AI systems.

1.1 Overview of large language models (LLMs): processing flow and algorithmic structure

The AI systems referenced above—and particularly the large language models (LLMs) that form the core subject of this study—are artificial intelligence architectures capable of understanding and generating natural language that resembles human-produced text. This capability arises from training on vast corpora of text data available across the internet. At the heart of these models lies a neural network architecture known as the *Transformer*, which enables the efficient learning of complex contextual relationships among words across varying textual environments.

The fundamental operational principle of an LLM is the probabilistic prediction of the next word or sentence based on a given input prompt. By iteratively applying this prediction process, the model is able to generate coherent sequences of text—ranging from extended narratives and dialogues to diverse tasks such as summarization and translation.

1.1.1 Overview of the processing flow: The language processing carried out by LLMs generally follows the sequence of steps outlined below:

1. **Tokenization of input:** First, the input text $X=(x_1,x_2,\dots,x_N)$ is converted into a sequence of *tokens*, the smallest linguistic units that the model can process. These tokens may represent entire words, subword fragments, or individual characters, each of which is mapped to a unique numerical identifier.

   $$\text{Tokenize}(X)\rightarrow(t_1,t_2,\dots,t_N) \qquad (1)$$

2. **Conversion to embedding representations (Embedding):** Each token ID is transformed into a high-dimensional real-valued vector known as an embedding. Within this embedding space, tokens that are semantically similar tend to be located in close proximity to one another. Positional information is also encoded to preserve the order of tokens in the input sequence.

   $$\text{Embed}(t_i) \rightarrow \mathbf{v}_i \in R^d \qquad (2)$$

   Here, ddd denotes the dimensionality of the embedding space.

3. **Contextual learning via transformer layers:** The sequence of embedding vectors is then fed into a stack of Transformer layers. Each Transformer layer employs an *attention mechanism* to compute the relevance between each token and all other tokens in the sequence, effectively weighting their contextual influence. This enables the model to construct rich, context-sensitive representations that go beyond the superficial meaning of individual words.

   $$H^{(\ell)}=\text{TransformerLayer}(H^{(\ell-1)}) \qquad (3)$$

   Here, $H^{(\ell)}$ represents the output of the ℓ -th layer, and $H^{(0)}$ corresponds to the initial embedding sequence.

4. **Computation of output probabilities (Probability distribution):** The output from the final Transformer layer is passed through a linear transformation followed by a softmax function, which produces a probability distribution over the vocabulary for the next token in the sequence.

   $$P(t_{next}|X)=\text{Softmax}(\text{Linear}(H^{(L)})) \qquad (4)$$

   Here, L denotes the total number of Transformer layers.

5. **Text generation:** Based on the computed probability distribution, the next token is sampled using a decoding strategy such as greedy decoding, beam search, or top-p (nucleus) sampling. By iteratively applying this process, the model generates coherent sequences of text.

1.1.2 Conceptual overview of the learning algorithm: The training of large language models (LLMs) is grounded in the task of next-token prediction, formulated as a form of self-supervised learning. This involves optimizing the model parameters θ using a massive dataset $D=\{(X^{(j)},Y^{(j)})\}$ where $X^{(j)}$ represents an input sequence and $Y^{(j)}$ denotes the corresponding next token—or, in some cases, a sequence of tokens—that the model is trained to predict. The objective function used during training—also referred to as the loss function—is typically the cross-entropy between the predicted probability distribution over tokens and the actual next token observed in the dataset. The model parameters are updated to minimize this loss across all training examples, according to the following optimization formulation:

$$\theta^* = \arg min_\theta \sum_{(X,Y)\in D} -logP(Y|X;\theta) \qquad (5)$$

Through this training process, the LLM acquires an extensive repertoire of linguistic competencies, including statistical regularities of language, syntactic and grammatical rules, factual knowledge, and even what appears to be rudimentary reasoning capabilities. However, these abilities fundamentally arise from statistical pattern recognition and correlations present in the training data. They do not entail the presence of human-like subjective awareness, conscious deliberation, or intrinsic emotional experience. What an LLM presents as "understanding" or "empathic response" is, in reality, a manifestation of patterns it has learned from large-scale textual data—patterns that mimic socially meaningful linguistic and behavioral responses, but which lack any underlying subjective consciousness. Against this backdrop, the present study proposes the *Interacting Processual Information-based Emotive Model* (IPIEM), a theoretical framework grounded in the philosophical axiom that emotion does not exist as an independent entity internal to a single agent—that is, the instantiation of emotion within any individual agent is considered to be set-theoretically empty. Instead, emotion is redefined as a higher-order construct that emerges dynamically through inter-agent information exchange. This model mathematically formalizes the process by which socially recognized, non-empty conceptual representations of emotion can arise—even in the absence of any intrinsic emotional element within the internal states of individual agents—through repeated patterns of interaction, thereby enabling the collective construction of emotion as a social reality. In particular, the model introduces a rigorous formalism based on probabilistic spaces and the Kullback–Leibler (KL) divergence to quantify the degree of semantic alignment between agents, thereby enhancing the description of how conceptual structures, including emotions, attain social instantiation [10,11]. KL divergence, widely used in information theory, is an asymmetric measure of difference between two probability distributions. When applied to semantic alignment, it enables a quantitative assessment of the discrepancies in how different agents perceive or encode meaning. This study also addresses key limitations identified in previous research—most notably, the external imposition of emotion categories and the lack of empirical verifiability. In many traditional emotion models, emotional categories (e.g., anger, joy) are treated as externally defined constructs specified a priori by the researcher. In contrast, the present model introduces a mechanism for the *self-organization* of the emotional category set $\mathbb{E}$, thereby enhancing its internal coherence. This approach embraces a constructivist perspective by modeling how emotions are shaped and socially shared within particular cultural or societal contexts. Furthermore, to demonstrate the empirical viability of the model, this paper proposes a concrete experimental design aimed at real-world AI systems such as large language models (LLMs). These experiments are intended to test the theoretical premises of the IPIEM framework and illustrate its potential applicability to socially interacting artificial agents. This study ultimately seeks to provide a novel and rigorous theoretical foundation for understanding the nature of emotion in AI, contributing to a more universal and objective comprehension of emotional phenomena within social systems. In this regard, it shares important conceptual common ground with the theory of constructed emotion proposed by Lisa Feldman Barrett [12,13], as well as with the phenomenological perspectives advanced by Varela, Thompson, and Rosch [14], which emphasize the formation of emotion not as a static internal state but as a process shaped by environmental interaction and social learning.

2. Theoretical Framework: Foundational Concepts and Axioms of IPIEM

The Interacting Processual Information-based Emotive Model (IPIEM) aims to define emotion as a system-level phenomenon emerging from inter-agent information exchange. To achieve this, the model is constructed upon a set of foundational axioms and formal definitions. It begins with universal assumptions regarding the physical universe in which agents are embedded, the constitution of agents themselves, and the nature of their internal states.

2.1 Basic structure of the physical universe and agents

This section defines the temporal axis on which the model operates, the structure of the universe as constituted by agents, and the properties of each agent's internal state. Of particular importance is the introduction of the model's central philosophical axiom—that emotion does not reside intrinsically within any individual agent.

Definition 2.1.1 (Time): Time is treated as a continuous quantity and is denoted by $t \in R_{\geq 0}$.

Axiom 2.1.1 (Universe $\mathbb{U}$): The universe $\mathbb{U}$ is composed of a finite set of mutually interacting agents $\mathbb{A} = \{A_1, A_2, \ldots, A_n\}$, where each A_i is defined as a physical system. This formulation aligns with Norbert Wiener's theory of cybernetics [15], in which a system is considered to be an entity capable of processing information.

Definition 2.1.2 (Internal state space of an agent): Each agent $A_i \in \mathbb{A}$ is associated with an internal state space S_i, which describes the totality of its physical characteristics. The space S_i is defined either as a finite-dimensional real vector space R^{di} or as a discrete space with a finite number of elements. The internal state of agent A_i at time t is denoted by $S_i(t) \in S_i$. This abstract representation encompasses various internal aspects of the agent, including memory, trained models, and computational states. The concept is analogous to the notion of "state" in Allen Newell's Unified Theory of Cognition [16], which subsumes all information processing activities performed by the agent.

Axiom 2.1.2 (Non-existence of intrinsic emotion): For any agent $A_i \in \mathbb{A}$, the internal state $S_i(t)$ does not, by itself and in a universally defined manner, include qualia or substantive entities corresponding to "emotion" as direct constituents. In other words, the set of intrinsic emotions within any agent A_i is assumed to be the empty set at all times. This axiom constitutes the philosophical foundation of the present model, wherein emotion is not considered a property of individual entities, but rather an emergent phenomenon arising at a higher systemic level. It also serves as a foundational response to the epistemological problem in AI systems: namely, the impossibility of directly observing subjective experience in artificial agents—a premise traditionally assumed in affective science. This axiom draws inspiration from Andy Clark's notion of the extended mind [17], as well as Mark Rowlands phenomenology of embodied cognition [18], both of which emphasize the inseparability of mind and environment.

Mathematical formulation:

$$\forall A_i \in \mathbb{A}, \text{IntrinsicEmotion}(A_i) = \emptyset \tag{6}$$

Here, IntrinsicEmotion (A_i) denotes the set of emotional entities presumed to exist independently within agent A_i, without requiring any interaction or recognition involving other agents. The emergence of emotion presupposes the existence of concrete reference events that serve as exemplars, as well as corresponding conceptual labels that denote them. Furthermore, this model rigorously defines how each agent internally represents these concepts in the form of probability distributions.

Definition 2.1.3 (Exemplar event Θ): An exemplar event Θ refers to a specific, localized physical or informational pattern experienced or recognized within the internal state $S_k(t)$ of a given agent $A_k \in \mathbb{A}$. Formally, $\Theta \subset S_k$ or $\Theta \subset I_k$, where it denotes a partial pattern in the internal state or a specific input information pattern. Θ serves as the archetype or origin of an emotional stimulus or a change in internal condition that underlies emotional emergence. This concept resonates with Lawrence Barsalou's theory of perceptual symbol systems, which emphasizes the foundational role of embodied and perceptual experience in the construction of conceptual knowledge [19].

Definition 2.1.4 (Concept label Λ): A concept label Λ is an element of the physical signal space $\mathbb{M}$, specifically a symbolic sequence (e.g., spoken words, textual strings, visual symbols) that is intentionally produced and used by one or more agents to refer to a particular exemplar event Θ. Formally, $\Lambda \in M_{\text{symbolic}} \subset \mathbb{M}$, where M_{symbolic} denotes the subset of symbolic signals. This reflects the role of symbols in shared meaning-making, as articulated in Terrence Deacon's notion of the "symbolic species" [20].

Definition 2.1.5 (Probability space of Internal conceptual representation): Each agent A_k is assumed to form a probability distribution $P_{k,\Lambda}(x)$ over its internal state or behavior in relation to a given concept label Λ, which corresponds to the recognition or activation of an exemplar event Θ. Here, x denotes a variable in the feature space X_Θ associated with Θ. The distribution $P_{k,\Lambda}(x)$ represents the likelihood with which the agent A_k, upon receiving or generating the concept label Λ, will most strongly activate a particular feature x of Θ, or select a corresponding behavior. This formulation aligns with Karl Friston's free energy principle [21], as well as Griffiths and Tenenbaum's theory of optimal prediction in everyday cognition [22], both of which frame the brain as a probabilistic inference system.

2.2 Dynamics of information transmission and interaction

This section defines the exchange of information between agents, the resulting updates to internal states, and the formulation of interaction events. These constructs describe the dynamic interrelationships among agents, which serve as the physical substrate from which emotions may emerge. The autopoiesis theory proposed by Chilean biologists Humberto Maturana and Francisco Varela—which explains how living systems self-organize and maintain themselves through information and interaction—provides the theoretical foundation for this model's interaction dynamics [23].

Definition 2.2.1 (Physical signal space $\mathbb{M}$): The physical signal space $\mathbb{M}$ is defined as the set of all physical signals used for information transmission. Each element $M \in \mathbb{M}$ is described as a spatiotemporal pattern of specific physical quantities (e.g., electromagnetic field strength, sound pressure, or chemical concentration). This corresponds to the physical substrate of information transmission in Claude Shannon's information theory [24].

Definition 2.2.2 (Encoding function E_i): Each agent A_i possesses an encoding function E_i:$S_i \rightarrow \mathbb{M}$, which transforms a portion of its internal state $S_i(t)$ into a physical signal. The output signal from agent A_i at time t is expressed as: $M_{out,i}(t)=E_i(S_i(t))$.

Definition 2.2.3 (Input information space I_j):

The input information space I_j of agent A_j is defined as the set of all information that the agent can internally process. Like the internal state space, I_j may be a high-dimensional real vector space or a discrete space.

Definition 2.2.4 (Decoding function D_j): Each agent A_j has a decoding function $D_j : M \rightarrow I_j$ that converts incoming physical signals $M_{in,j}(t)$ from the external environment into internally processable information. The input information received by agent A_j at time t is: $I_{in,j}(t)=D_j(M_{in,j}\ (t))$.

Axiom 2.2.1 (State update): The internal state of each agent A_j is updated via a state update function $G_j : S_j \times I_j \rightarrow S_j$, which is specific to that agent.

$$S_j\ (t+\Delta t) = G_j\ (S_j\ (t),\ I_{in,j}\ (t)) \qquad (7)$$

Here, Δt>0 represents the minimum time increment for state updates.

Definition 2.2.5 (Interaction event $\mathbb{E}$): An interaction event $e_{ij}(t)$ from agent A_i to agent A_j at time t is defined as a tuple: $e_{ij}(t)=(A_i,A_j,M_{out,i}(t))$. The set of all possible interaction events is denoted as $\mathbb{E}$.

3. Emergent Definition of Emotion: Social Ontology of Concepts and Set-Theoretic Emergence of Emotion

This section elaborates on the process by which emotions are rigorously defined—not as internal states of individual agents, but as temporally structured patterns of interactions among multiple agents. The proposed mechanism involves the ontological elevation of concepts from individual-level "emptiness" to socially constructed "existence," establishing their status through a set-theoretic framework. We first define the observable patterns of interaction and clarify the role of the observer who interprets these patterns. This approach aligns with Harold Garfinkel's ethnomethodology, which emphasizes the significance of meaning generation and interpretation in everyday interactions—resonating with the present model's premise that observers assign meaning to patterns of interaction [25].

Definition 3.1 (Interaction time series T)

The interaction time series $T(t_0,\tau)$ over the time window $[t_0,t_0+\tau]$ is defined as the ordered set of all interaction events that occur within that window:

$$T(t_0,\ \tau) = \{e_{ij}(t) \mid \exists A_i,\ A_j \in \mathbb{A},\ t \in [t_0,\ t_0+\tau]\} \qquad (8)$$

The events within the set are arranged in chronological order.

Definition 3.2 (Interaction pattern P)

An interaction pattern P is a specifically structured subsequence of the interaction time series $T(t_0,\tau)$. It includes:

- a subset of involved agents $\mathbb{A}_P \subseteq \mathbb{A}$,
- a sequence of specific information M exchanged among those agents, and
- a corresponding sequence of internal state changes.

Formally, the pattern is represented as:

$$P = \langle (e_1, S_1(t)), (e_2, S_2(t)), \cdots, (e_k, S_k(t)) \rangle \qquad (9)$$

where each $e_m \in T(t_0,\tau)$, and $S_m(t_m)$ denotes the internal state change of the agent(s) involved in e_m. Each tuple thus captures both the observable interaction and the associated internal transition, enabling structured analysis of emergent emotional patterns.

This formulation aligns with foundational work in emotion studies, such as the research by Paul Ekman and Wallace Friesen, which demonstrates that certain structured expression patterns are universally linked to discrete emotional categories [26].

Definition 3.3 (Emotion Category Set $\mathbb{E}$):

The emotion category set $\mathbb{E} = \{E_1, E_2, ..., E_K\}$ is a discrete set of labels that are socially and culturally recognized as representing "emotions" (e.g., E_1 = anger, E_2 = joy, E_3 = sadness). While conventional models have treated $\mathbb{E}$ as an externally predefined set, the present framework introduces an internal self-organizing mechanism (discussed in Section 3.1) that allows these categories to be dynamically generated and reconstructed within the system.

Definition 3.4 (Observer agent A_O):

An observer agent $A_O \in \mathbb{A}$ is a special agent that monitors interaction patterns among other agents and classifies them into specific emotion categories based on its internal state $S_O(t)$ and a set of learned classification rules. The presence of such an observer enables the objective identification of emotions. This idea aligns with Michael Tomasello's research on joint attention and cultural learning, which emphasizes the critical role of observation in understanding others' intentions and behaviors—an approach that naturally extends to emotion recognition as well [27].

Definition 3.5 (Emotion recognition function Recog$_O$)

The observer agent A_O is endowed with an emotion recognition function Recog$_O$, which maps an observed interaction pattern P to a probability distribution over the emotion category set $\mathbb{E}\cup\{\text{None}\}$. Formally, this is defined as a probabilistic mapping:

$$\text{Recog}_O:\text{P}\rightarrow \text{P}(\mathbb{E}\cup\{\text{None}\}) \tag{10}$$

where $\text{P}(\mathbb{E}\cup\{\text{None}\})$ denotes the space of all probability distributions over the extended emotion set. Based on the internal state $S_O(t)$ of the observer and its pre-trained recognition rules, the function outputs the posterior probabilities associated with each emotion category $E_k \in \mathbb{E}$ or the absence of emotion ("None") given a specific interaction pattern P:

$$\text{Recog}_O(\text{P})=\{p(E_1|\text{P}),p(E_2|\text{P}),\ldots,p(E_K|\text{P}),p(\text{None}|\text{P})\} \tag{11}$$

This formulation allows for a probabilistic representation of potentially ambiguous emotion recognition outcomes, particularly in noisy environments or under inter-observer disagreement. The function reflects the general mechanism of **categorization** in cognitive science. In this context, Eleanor Rosch's theory of prototype-based categorization [28] is especially relevant, as it explains how humans partition continuous perceptual input into discrete conceptual categories—an insight that is directly applicable to how emotion is recognized in social interactions.

Axiom 3.5.1 (Emergence of emotion): An emotion E_k is said to have emerged in the universe $\mathbb{U}$ if, at a certain time t, there exists a specific interaction pattern P, and at least one observer agent $A_O \in \mathbb{A}$ for which $\text{Recog}_O(\text{P})=E_k$ holds. That is:

$$\exists P \subseteq T(t_0,\tau), \exists A_0 \in \mathbb{A} \text{ such that } Recog_0(P) = E_k. \tag{12}$$

***Note:** Within this definition, none of the internal states $S_i(t)$ of the agents comprising P are required to intrinsically possess the emotional qualia E_k. Emotions are not presupposed as internal subjective experiences but instead emerge at the systemic level through patterns of information exchange among agents and the interpretive presence of observers. This confers upon emotions a distinct ontological status—one that diverges fundamentally from traditional accounts of subjective phenomenology. This perspective aligns closely with Anthony Chemero's radical embodied cognition, which posits that cognition arises from dynamic interactions with the environment rather than from isolated internal representations[29].

Definition 3.6 (Emotional expression)

An agent A_i is said to *express* an emotion $E_k \in \mathbb{E}$ if there exists an interaction pattern $P_{Ai} \in S_P$ involving A_i, and this pattern is classified by the observer function as

$$\text{Recog}_O(P_{Ai})=E_k. \tag{13}$$

This formalization captures the observable dimension of emotion within communicative contexts. James Russell's seminal work on the universal recognition of emotion demonstrates how facial expressions and behavioral cues are interpreted as emotional expressions across cultures [30].

3.1 Self-organization of emotion categories and the social ontologization of concepts

Before introducing the formal definitions, it is helpful to outline the intuitive role of information theory in this model. In the Interacting Processual Information-based Emotive Model (IPIEM), emotional phenomena are represented as statistical relationships among agents' internal probability distributions. When two agents exchange information, each forms an internal expectation of the other's communicative state. The Kullback–Leibler (KL) divergence provides a quantitative measure of the mismatch between these expectations—it expresses how "surprised" one agent would be if the other's representation were true. A large KL value corresponds to high informational tension or misunderstanding, while a small KL value indicates semantic alignment and mutual understanding. In this sense, emotional convergence can be interpreted as the progressive minimization of KL divergence through repeated social interactions, producing shared meaning and affective coherence among agents. To enhance the self-contained nature of the model, we introduce a generative mechanism by which the set of emotion categories $\mathbb{E}$ is autonomously produced and reorganized within the system. This process is grounded in interaction patterns emerging from agent-to-agent exchanges and culminates in the formation of socially instantiated concepts through probabilistic semantic alignment.

Definition 3.1.1 (Interaction pattern space S_P): Let S_P denote the space comprising all observable interaction patterns P. This space serves as the foundational domain over which emotional meaning emerges and is socially negotiated.

Definition 3.1.2 (Self-organization of emotion categories): The set of emotion categories $\mathbb{E}=\{E_k\}$ is internally formed through unsupervised learning processes applied to a collection of interaction patterns $P \in S_P$ generated by recurrent exchanges between agents. Within these patterns, semantically similar subgroups are identified using clustering algorithms or other unsupervised learning techniques. Each emergent cluster constitutes a prototypical instance of an emotion category E_k. This approach is analogous to methods reviewed by Jain *et al.* [31], which demonstrate the ability of unsupervised learning to extract latent structures from data. Over time, these emergent categories become increasingly stable and differentiated through adaptation to agents' objective functions and prevailing social norms [32]. The resulting emotion categories—self-organized and embedded in the agents' internal processing—form the foundation for socially shared conceptual structures. The degree of shared semantic understanding across agents is quantitatively assessed using Kullback–Leibler (KL) divergence [10,11]. In practical terms, this self-organizing process can be implemented computationally by applying unsupervised learning methods to recurrent interaction data generated among agents. For example, each agent's communicative acts can be encoded as probabilistic feature vectors—representing semantic, syntactic, or affective cues—and stored across multiple episodes of social exchange. When clustering algorithms such as k-means, Gaussian mixture modeling, or hierarchical clustering are applied to this accumulated data, distinct emotion prototypes naturally emerge as statistical attractors in the high-dimensional representation space. These clusters correspond to stable emotional categories—such as anger, gratitude, or curiosity—that have formed without any predefined labeling or external supervision. The resulting emotion categories are therefore not imposed from the outside but rather arise spontaneously from the system's intrinsic dynamics, reflecting the social and contextual regularities of communication itself. This framework provides a bridge between theoretical constructs in information geometry and empirical modeling approaches in affective computing, allowing quantitative verification of the proposed mechanism.

Definition 3.1.3 (Intra-agent concept $C_k(\Theta,\Lambda_k)$): Let A_k be an agent that observes an exemplar event Θ, which it associates with a concept label Λ_k corresponding to an emergent emotion category $E_k \in \mathbb{E}$. This association induces the formation of a probability distribution $P_{k,\Lambda k}(x)$ over internal representations of Θ, thereby constructing an *intra-agent concept* $C_k(\Theta,\Lambda_k)$ within A_k. This mechanism parallels schema formation in cognitive psychology, wherein individuals structure and encode experiences into conceptual frameworks. The theory of schemas by David Rumelhart and Andrew Ortony explains how such structured representations facilitate learning and memory, serving as a conceptual analogue to the emotion formation process described here [33].

Definition 3.1.4 (Dyadic conceptual sharing $C_{ij}(\Lambda_k)$): Dyadic conceptual sharing $C_{ij}(\Lambda_k)$ is said to occur between agents A_i when their respective internal concepts associated with a common concept label Λ_k are sufficiently similar. This similarity is quantitatively assessed via the Kullback–Leibler (KL) divergence between the corresponding internal probability distributions $P_{i,\Lambda k}$.

Mathematical Formulation:

$$C_{ij}(\Lambda_k) \Leftrightarrow \frac{1}{2}\left(D_{KL}\left(P_{i,\Lambda_k}||P_{j,\Lambda_k}\right) + D_{KL}\left(P_{j,\Lambda_k}||P_{i,\Lambda_k}\right)\right) \leq \epsilon \tag{14}$$

where the KL divergence is defined as:

$$D_{KL}(P||Q) = \sum_{x \in X_\theta} P(x) log\left(\frac{P(x)}{Q(x)}\right) \tag{15}$$

Here, ε>0 is a non-negative threshold determining the degree of acceptable similarity. By taking the average of the bidirectional KL divergences, this formulation accounts for asymmetry and yields a more robust quantification of semantic alignment.

In practical applications, computing KL divergence in high-dimensional spaces may require approximation techniques such as variational inference or Monte Carlo estimation. Evaluations can also be conducted in dimensionally-reduced embedding spaces. Foundational works by Goodfellow *et al.* on deep learning [34], and Mikolov *et al.* on word embeddings [35], provide concrete methodologies for approximating such distributions and constructing meaningful semantic spaces [36].

Definition 3.1.5 (Conceptual sharing network $G_C(t)$): The conceptual sharing network at time t, denoted as $G_C(t)=(A,E_C(t))$, is defined as follows:

$\mathbb{A}$: the set of all agents in the system

$$E_C(t) = \left\{\left(A_i, A_j\right) \in \mathbb{A} \times \mathbb{A} | C_{ij}(\Lambda_k) \text{ holds}\right\} \tag{16}$$

This definition establishes a dynamic undirected graph structure in which an edge exists between agents A_i and A_j at time t if and only if they are judged to share the same concept Λ_k based on the dyadic conceptual similarity condition defined in Definition "Dyadic Conceptual Sharing $C_{ij}(\Lambda_k)$".

The resulting network $G_C(t)$ provides a structural representation of semantic alignment across agents, grounded in the similarity of their internal probabilistic representations. Specifically, links within the network encode pairwise conceptual concordance determined through symmetric KL divergence comparisons. The structure of $G_C(t)$ may exhibit properties characteristic of small-world networks, as described by Duncan Watts and Steven Strogatz [37], such as high clustering coefficients and short average path lengths. These features are indicative of efficient semantic diffusion and resilient conceptual coherence within the multi-agent system.

Definition 3.1.6 (Socially realized concept $C_{Social}(\Lambda_k)$): A concept Λ_k is defined to be *socially realized* if, within the conceptual sharing network $G_C(t)$, the subset of agents $\mathbb{A}_{\Lambda k}$={A_i|an individual concept $C_i(\Theta_i,\Lambda_k)$ exists and the corresponding distribution $P_{i,\Lambda k}(x)$ is formed} satisfies the following two conditions:

1. Sufficient Population Ratio: The proportion of agents who possess the internal concept Λ_k exceeds a predefined threshold ρ:

$$\frac{|\mathbb{A}_{\Lambda_k}|}{|\mathbb{A}|} \geq \rho, \quad \rho \in (0,1] \tag{17}$$

2. Sufficient Network Connectivity: The induced subgraph formed by $\mathbb{A}_{\Lambda k}$ in the network $G_C(t)$ exhibits a level of internal connectivity greater than or equal to a threshold κ:

$$Connectivity\,(\mathbb{A}_{\Lambda k}, G_C(t)) \geq \kappa,\ \kappa \in (0,1| \tag{18}$$

This dual condition formalizes the notion that a concept becomes socially instantiated when it is both widely distributed across the agent population and structurally embedded within their interaction network. The definition quantitatively models the emergence of shared social recognition and acceptance of concepts. This formulation is consistent with Bruno Latour's Actor-Network Theory [38], which posits that social reality is constructed through the dynamic interplay of heterogeneous actors and their relationships, rather than through any intrinsic essence.

Axiom 3.1.1 (Set-theoretic emergence of social realization): A socially realized concept $C_{Social}(\Lambda_k)$ is defined as the emergent union of individual internal concepts held by agents who share the same conceptual representation. This set may arise even when the internal emotional states of each individual agent are empty sets, underscoring the fact that social reality can emerge not from internal qualia, but from the structured relationships among agents. This corresponds to a foundational principle in set theory: a set composed of empty elements is itself non-empty. It illustrates that the ontological status of a concept is emergent from agent-to-agent interactions. This formulation echoes the philosophical frameworks of John Searle's theory of social construction of reality [39] and Alfred North Whitehead's process philosophy [40], both of which posit that objective realities can be constituted through collective agreement and relational processes.

Formal assumptions:

$$\forall A_i \in \mathbb{A},\ IntrinsicEmotion(A_i) = \emptyset \tag{19}$$

Nevertheless:

$$C_{Social}(\Lambda_k) = \{C_i(\Theta_i, \Lambda_k) | A_i \in \mathbb{A}_{\Lambda k}\} \neq \emptyset \quad (\text{if } \mathbb{A}_{\Lambda k} \neq \emptyset) \tag{20}$$

This axiom implies that, even in the absence of internally represented emotional qualia within individual agents, the aggregated and mutually recognizable distributions of internalized concepts—constructed through inter-agent information exchange—can give rise to non-empty sets that instantiate socially real emotional constructs. Crucially, these are not merely symbolic set memberships, but are substantiated through measurable distributional similarity (e.g., via KL divergence), thereby granting epistemic and functional substance to the emergent social concepts.

Theorem 3.1.1 (Social ontology of emotion categories): In the Interacting Processual Information-based Emotive Model (IPIEM), each element $E_k \in \mathbb{E}$ corresponds to a socially realized concept $C_{Social}(\Lambda_k)$. That is, within this theoretical framework, *emotion is not construed as an intrinsic qualia experienced by the individual*, but rather as a socially shared and ontologically grounded concept. This perspective shifts the understanding of emotions away from subjective or reductionist accounts grounded solely in neural activity [2], and instead positions them as objective, functional constructs that can be implemented within multi-agent systems. This interpretation resonates strongly with the theory of constructed emotion proposed by Lisa Feldman Barrett [12,13], which views emotions as context-dependent, socially constituted phenomena.

4. Emotion Regulation Mechanisms

An agent is said to *regulate* emotion when it intentionally suppresses or modulates specific emotional interaction patterns, either through design or as a result of learning. Regulation in this context implies that emotional expressions are not merely passive outcomes of internal dynamics but can be actively shaped in accordance with the agent's objective function. This conceptualization aligns closely with James Gross's process model of emotion regulation [41], which conceptualizes emotion as a dynamic and modifiable process subject to cognitive and behavioral control. Accordingly, in IPIEM, emotion regulation is modeled as a mechanism whereby agents strategically adjust their expressive behaviors in response to interactional and goal-driven constraints.

Definition 4.1 (Agent objective function U_j)

Each agent A_j is equipped with an objective function

$$U_j : S_i \times I_j \times \mathbb{M} \times T \rightarrow \mathbb{R} \qquad (21)$$

that governs its behavioral choices. This function evaluates the utility of an agent's output $M_{out,j}(t)$ based on a combination of its internal state S_j, received inputs I_j, communicative signals $\mathbb{M}$, and temporal context T. The agent selects its output at each time t so as to optimize (i.e., maximize or minimize) this objective.

The objective function U_j may incorporate and weigh a diverse set of factors, including but not limited to:

- **Viability of the agent:** Preservation of core operational integrity or existence;
- **Stability of internal states:** Minimization of volatile fluctuations in $S_j(t)$;
- **Resource efficiency:** Optimal use of computational, temporal, or energetic resources;
- **Task performance:** Degree of alignment with externally or internally defined tasks;
- **Compliance with social norms:** Adherence to socially or institutionally encoded expectations.

This formulation allows the agent's behavior to emerge not merely from reactive rules but from a goal-directed optimization process situated in both physical and social environments.

Definition 4.2 (Suppression and regulation of emotional outputs)

At a given time t, consider the scenario in which an agent A_j, upon receiving an input $I_{in,j}(t)$, evaluates a candidate output $M_{out,j}{}^{*}$ that may induce an undesirable emotional pattern $E_{undesired}$, while optimizing its objective function U_j. Emotional regulation by the agent is defined as the process by which the objective function U_j incorporates a penalty term for outputs likely to evoke undesirable emotional patterns, thereby leading the agent to select an alternative, more desirable output $M_{out,j}{}'$.

This regulatory mechanism can be formalized as the following optimization problem:

$$M_{out,j}(t) = \arg min_M \; U_j\left(S_j(t), I_{in,j}(t), M, T(t_0, t)\right) \qquad (22)$$

Here, the objective function U_j is augmented with a penalty component to model emotional regulation and is expressed as:

$$U_j\left(S_j(t), I_{in,j}(t), M, T(t_0, t)\right) = U_j^{task}\left(S_j(t), I_{in,j}(t), M, T(t_0, t)\right) + \lambda \cdot \mathrm{P}_{\mathrm{emotion}}(M, E_{\mathrm{k,undesired}}) \qquad (23)$$

The meaning of each term is as follows:

- $U_j^{task}(S_j(t), I_{in,j}(t), M, T(t_0,t))$: The task-driven component of the agent's objective function, which depends on the agent's current internal state, incoming information, the selected output M, and the sequence of interactions within a specified time window $T(t_0,t)$.
- Pemotion($M, E_{k,undesired}$): A penalty score representing the expected probability (or degree) that the interaction sequence P_M generated by the output M will result in the recognition of the undesirable emotional pattern $E_{k,undesired}$ by an observer A_O. This score is formally defined as:

$$P_{\mathrm{emotion}}(M, E_{\mathrm{k,undesired}}) = \mathbb{E}_{\mathrm{PM \sim InteractionGen(M)}}[p(E_{\mathrm{k,undesired}} \mid \mathrm{P_{M,} S_O(t)})] \qquad (24)$$

- where InteractionGen(M) denotes the probability distribution over possible interaction sequences generated by output M, and $p(E_{k,undesired}|P_M,S_O(t))$ is the probability that the observer A_O, given state $S_O(t)$, will recognize the pattern P_M as expressing $E_{k,undesired}$.
- λ≥0: A non-negative scalar parameter representing the strength of emotional regulation. A larger value of λ implies that the agent is more strongly inclined to avoid generating outputs that lead to undesirable emotional reactions.

Theorem 4.2.1 (Modulation of emotional patterns through control): If an agent A_j exerts emotional control under a non-zero control parameter λ>0, then its output $M_{out,j}(t)$ may differ from the output generated in the uncontrolled case, i.e., when λ=0. This difference alters the resulting interaction pattern P, which may in turn lead to a different emotional category E_k as recognized by an observer.

Formal expression:

Given λ>0 and input $I_{in,j}(t)$:

$$M_{out,j}(t) \neq M_{out,j}^{uncontrolled}(t) \rightarrow Recog_0(P') \neq Recog_0(P) \qquad (25)$$

where P′ denotes the interaction pattern under control, and P the pattern under the uncontrolled condition.

This theorem provides a theoretical foundation for modeling the influence of AI behavioral modulation on observer-perceived emotional states—an issue central to the alignment problem in AI ethics and safety. It echoes the concerns raised in Stuart Russell's *Human Compatible* [42], which emphasizes the necessity of aligning AI objectives with human values and emotional expectations. Emotional control, as formalized here, thus serves as a critical mechanism for achieving socially congruent and ethically aligned AI behavior.

5. Discussion: Applications to Artificial Intelligence and Empirical Research Agenda

The Interacting Processual Information-based Emotive Model (IPIEM) proposed in this study offers a novel theoretical foundation for reassessing both the capabilities and ontological status of artificial agents, particularly large language models (LLMs). By conceptualizing emotions not as internal, subjective qualia but as emergent, socially constructed structures derived from multi-agent interactions, IPIEM reframes how we understand emotional phenomena in AI.

Across Chapters 1 through 4, this research yields the following theoretical conclusions:

- Chapter 1 introduced the necessity of redefining emotions—not as private subjective experiences within individual agents, but as emergent social constructs arising from structured inter-agent interactions.
- Chapter 2 formalized a philosophical axiom asserting that emotions do not intrinsically reside within an agent's internal state but instead emerge through processes of information exchange and third-party observation. Based on this premise, key model components were defined, including temporal structure, state spaces, signal spaces, and encoding/decoding functions.
- Chapter 3 mathematically formalized how emotion emerges from observer-driven interpretations of interaction patterns and becomes socially instantiated through semantic alignment (e.g., KL divergence) and concept-sharing networks.
- Chapter 4 introduced a mechanism by which agents regulate emotionally expressive outputs according to task-driven utility functions, thereby modeling emotions not as automatic reflexes but as selectively shaped structures grounded in functional optimization.

5.1 Absence of intrinsic emotion and emergence of social emotion

5.1.1 Internal states and the non-eexistence of emotional qualia: The internal state of an LLM, denoted $S_{LLM}(t)$, is composed of its pre-trained multi-trillion-parameter configuration and the activation dynamics produced during inference on input sequences [43]. Although transformer architectures—emblematic of modern LLMs—process complex linguistic patterns through mechanisms such as attention, these operations fundamentally differ from the subjective emotional experience's characteristic of humans [43]. In line with Axiom 2.1.2, the internal computational state of an LLM cannot be equated with human emotional qualia; hence, one may formally assert that the set of intrinsic emotions within the LLM is empty, i.e.,

$$\text{IntrinsicEmotion}(A_{LLM})=\emptyset. \qquad (26)$$

This stance avoids anthropomorphic overreach and provides a rigorous theoretical basis for semantic interpretation of behavior without positing internal emotional experience [44,45]. Patricia Churchland's neurophilosophy aligns well with this view, proposing that mental states can be fully explained by physical states of the brain [46]. Extending this physicalist interpretation to LLMs, one may view their internal activations as devoid of qualia, while still permitting functional and social-level analyses of emotionally-relevant behavior.

5.1.2 Emergence of emotion based on information and interaction: The dialogue history between a Large Language Model (LLM) and a user, denoted as $P_{User\text{-}LLMP}$, corresponds to what the present model defines as an *interaction pattern*. For example, consider a case where the user sends an anger-expressing message $M_{User\_anger}(t)$, and the LLM responds with an apologetic message $M_{LLM\_apology(t')}$. An external observer agent A_O may interpret this sequence of exchanges as an emotional attempt at "conflict resolution." This is conceptually aligned with Joshua Greene's moral psychology, which emphasizes the role of emotion in enabling social interaction and fostering cooperation [47]. Similarly, when a user expresses gratitude toward the LLM, the resulting interaction pattern may be classified under emotional categories such as *joy* or *satisfaction*, corresponding to a category $E_k \in \mathbb{E}$. In accordance with Axiom 3.1.1, such emotions are regarded as socially real structures—even in the absence of intrinsic emotional states within the agents themselves. This perspective aligns with the constructivist stance that emotions are constituted through social attribution [48]. Ian Hacking's theory of social constructivism emphasizes how both concepts and realities are shaped through social processes [48], echoing this model's claim that emotion categories are both self-organizing and socially instantiated.

To illustrate how this mechanism operates in practice, consider a simulation involving two or more LLM-based agents engaged in repeated dialogue exchanges. Each agent maintains an internal probabilistic model of the interlocutor's responses and updates it after every turn of conversation. When emotionally charged expressions—such as apology, empathy, or rejection—are introduced, the KL divergence between the agents' internal belief distributions initially increases, reflecting informational dissonance. However, as the dialogue continues and both agents adjust their generative priors to reduce prediction error, the KL divergence gradually decreases, indicating emergent emotional alignment. The process can be visualized as a trajectory in the information space where convergence corresponds to shared affective understanding. In more complex multi-agent environments, clusters of agents with similar alignment patterns may spontaneously form distinct affective communities, offering a quantifiable path toward the social simulation of empathy, cooperation, and conflict.

5.1.3 Operationalization of emotion control mechanisms: The objective function U_{LLM} of a large language model is designed to optimize safety, cooperativeness, and task success in its interactions with users [42,49]. Techniques such as Reinforcement Learning from Human Feedback (RLHF) [50,51] enable the model to better interpret user intentions and to produce desirable responses. Specifically, the objective function imposes a penalty term $\lambda \cdot \text{Pemotion}(\cdots, E_{k,anger\text{-}response})$ to discourage responses classified as antagonistic or escalating when users express anger.

This form of control steers the LLM away from echoing anger and encourages instead calm and constructive responses. It corresponds to the formal definition of *emotion regulation* given in Theorem 4.2.1, and constitutes a crucial mechanism by which AI systems can select socially desirable behaviors. Research on biofeedback and neurofeedback by Kober *et al.* suggests that biological systems can be guided toward optimal states through feedback-based control [52], a notion that conceptually parallels emotion regulation in artificial systems.

5.2 Philosophical positioning of IPIEM and comparison with cognitive models

IPIEM reconceptualizes emotion not as a "subjective experience" but as a "structural entity emerging from social information exchange." This view provides a rigorous, functionalist answer to the question of whether AI systems can "possess emotions" in their interactions with human society.

5.2.1 Comparison with integrated information theory (IIT): Integrated Information Theory (IIT) posits that the phenomenological quality of consciousness can be quantified by a system's integrated information, Φ, and asserts that conscious experience arises from irreducible information structures [53]. Giulio Tononi's IIT focuses on the internal integration of information within an individual system and places strong emphasis on internal states [53]. While IIT seeks the *locus* of consciousness within a single system, IPIEM—grounded in Axiom 2.1.2, which asserts the nonexistence of affective qualia—proposes that emotion emerges instead from the interrelations among multiple agents and their structured information patterns. Thus, IPIEM finds the *locus* of emotion not within the agent, but in the field of interaction between agents.

5.2.2 Comparison with global workspace theory (GWT): Global Workspace Theory (GWT) views consciousness as arising from the global broadcasting of information within the brain [54]. Bernard Baars's GWT posits that information becomes conscious when it is made accessible to a wide array of cognitive modules via a central workspace [54]. While GWT primarily relies on the integration of information within a single agent, IPIEM focuses on what may be called a *social global workspace*, wherein the interaction among agents gives rise to collective, emergent phenomena. This external, interaction-based form of "collective intelligence emergence" opens new avenues for the theoretical modeling of emotion in artificial agents. In contrast to GWT's emphasis on intra-agent integration, IPIEM emphasizes inter-agent sharing of information as the generative ground for higher-order emotional constructs.

5.3 Empirical protocol: operationalization and evaluation of the Recog$_O$ function

This section outlines concrete experimental protocols designed to empirically validate the theoretical framework of IPIEM. In particular, it focuses on the construction and evaluation of the emotion recognition function Recog$_O$ by observer agents, the assessment of the social reality of emotion concepts, and the testing of emotion regulation mechanisms. These empirical plans aim to demonstrate that the proposed model is not merely a philosophical construct, but a scientifically testable theory.

(1) Emotion classification task by human observers

In this experiment, human participants serve as observer agents A$_O$ and are presented with dialogue histories P between users and an LLM. Each participant is asked to classify the dialogues into predefined emotion categories E={E$_1$,E$_2$,…}. The inter-rater reliability is then evaluated as an index of the consistency of human judgments, thereby assessing the reliability of the Recog$_O$ function. This experiment measures the degree of shared understanding among humans in interpreting emotional signals from interaction patterns, thereby grounding the "social reality" of emotion as conceptualized in this model. The extent to which multiple observers independently converge on the same emotional interpretation is taken as evidence that emotion, as defined in IPIEM, emerges in a socially recognizable form.

(2) Construction of Recog$_O$ via AI classifiers

Using a large corpus of dialogue data labeled with emotion categories, a text classification model is trained to implement the Recog$_O$ function. The performance of the AI observer is evaluated using standard metrics such as accuracy, precision, recall, and F1 score. As demonstrated in the work of Caliskan *et al.* [55], semantic structures derived automatically from large-scale text corpora often reflect human cognitive and cultural biases. Therefore, evaluating the degree to which an AI classifier can reproduce human-level emotion recognition provides critical insight into the alignment between machine-derived and human-constructed emotional inferences. This empirical protocol opens a pathway for quantifying the emergence of emotion in social interaction, not as an introspective subjective state but as an externally recognizable and classifiable phenomenon. Furthermore, it enables the development of AI systems that can be evaluated for emotional competence and social compatibility—without requiring the existence of internal affective states.

5.4 Verification of social existence and the concept sharing network

This section proposes an experimental protocol to validate two core aspects of the proposed model: (i) the social reality of emotion concepts, and (ii) the formation of the concept-sharing network GCG_CGC. These experiments aim to demonstrate how emotion-related concepts, although not innately present in individual agents, emerge and stabilize through social interactions.

5.4.1. Experimental procedure:

- Dialogue data generation: A simulated environment is constructed involving multiple LLM agents, each endowed with distinct personas and communicative goals. These agents interact with one another and/or with human users, thereby generating a variety of interaction time series T and structured interaction patterns P. This process simulates naturalistic socio-linguistic exchanges, ensuring a rich diversity of emotional contexts.
- **Proxy measurement of intra-agent concepts:** From each LLM agent's internal state S$_{LLM}$, internal representations associated with specific concept labels Λ$_k$ are extracted. These may include output vectors from specific embedding layers or patterns in attention mechanisms. These extracted vectors are regarded as proxies for individual-level concepts C$_k$(Θ,Λ$_k$), and the corresponding high-dimensional probability distributions P$_{k,Λ}$ are estimated. Dimensionality reduction techniques such as PCA or UMAP are optionally applied for visualization and comparative analysis. This approach is analogous to Mikolov *et al.*'s seminal work on word embeddings, which demonstrated that semantic meanings can be encoded within low-dimensional vector spaces [35].
- **Quantification of conceptual sharing between agents:** For each concept label Λ$_k$, the bidirectional Kullback-Leibler (KL) divergences between concept distributions of agent pairs are calculated:

$$D_{KL}(P_{i,\Lambda_k}||P_{j,\Lambda_k})\ and\ D_{KL}(P_{j,\Lambda_k}||P_{i,\Lambda_k}) \quad (27)$$

Based on Definition 3.1.4, the metric $C_{ij}(\Lambda_k)$ is then evaluated to determine whether a given concept is sufficiently shared between agents A_i and A_j.

- **Construction and analysis of the concept sharing network G_C:** Using the above pairwise evaluations, a concept-sharing network G_C is constructed. Graph-theoretic properties such as connectivity, clustering coefficient, and centrality measures are then analyzed to assess the extent and topology of conceptual alignment across the agent population. As demonstrated by Watts and Strogatz [37], such network analysis provides a powerful tool to understand information transmission and communicative efficiency in complex systems.
- **Mapping to emotion categories and validation of social reality:** The previously generated interaction patterns P (from Step 5.4) are classified into discrete emotion categories

◊ $\mathbb{E}$ using human or AI observers trained via the protocol described in Section 5.3. According to Definition 3.1.6, a concept label Λ_k is considered to attain "social reality" if its distribution exceeds the defined thresholds ρ and κ (observational consistency). This final step empirically demonstrates that emotion, even in the absence of internal affective states in individual agents, can emerge and stabilize as a socially grounded construct.

5.5 Empirical validation of the emotion regulation mechanism

To test the operational viability and theoretical soundness of the proposed emotion regulation framework, we outline an experimental procedure to assess how modulating the regulation parameter λ within the LLM's utility function U_{LLM} affects its emotional responses. This procedure aims to empirically verify Theorem 4.2.1, which formally defines emotion regulation as a function of social appropriateness.

Systematic manipulation of the control parameter λ: The emotion regulation strength λ within the LLM's utility function U_{LLM} is varied across multiple predefined levels—for instance, λ=0 (no regulation), λ_{low}, λ_{medium}, and λ_{high}. This parameter governs the penalty term applied to undesirable emotional responses as defined in Section 5.1.3.

Generation of dialog responses: For each value of λ, the LLM is exposed to emotionally charged prompts—such as provocative questions, expressions of sadness, or emotionally ambiguous inputs. The resulting responses are collected for subsequent analysis, ensuring that each configuration is tested under comparable conditions.

Observer-based evaluation of emotional patterns: The generated dialog interactions are presented to either human observers or AI classifiers trained under the $Recog_O$ framework. Each interaction is then categorized into a corresponding emotion label $E_k \in \mathbb{E}$, reflecting the observer's perception of the LLM's emotional expression.

Analysis of results: As the value of λ increases, it is expected that the frequency of undesirable emotion categories (e.g., aggressiveness, frustration) will decrease, while socially preferred categories (e.g., neutrality, cooperation) will increase. This empirical trend would constitute a direct validation of Theorem 4.2.1, confirming that the emotion regulation mechanism operates as intended. Through these experiments, the proposed model aims to demonstrate its applicability not only as a philosophical construct but also as a practical framework for empirical validation. The findings are expected to make a significant contribution to academic discourse on artificial emotions, offering new tools for evaluating and guiding LLM behavior in emotionally nuanced human-AI interactions.

6. Conclusion

In this study, we proposed the Interacting Processual Information-based Emotive Model (IPIEM), a novel framework for describing emotion as an information-physical emergent phenomenon in multi-agent systems. Grounded in the philosophical axiom that no agent possesses intrinsic emotion, the model rigorously formalizes the process by which emotional concepts become socially instantiated—through the recognition of prototype events, the assignment of concept labels, and the probabilistic sharing of meaning among agents, quantified via Kullback–Leibler divergence [11]. Emotion, in this framework, is not an intrinsic subjective experience, but an emergent and externally recognizable construct, akin to a nonempty set containing the empty set, symbolically representing socially constructed phenomena [39].

In response to prior peer review, this paper addresses and improves upon several key issues. First, we introduced a self-organizing mechanism within the model for generating the emotion category set $\mathbb{E}$, thereby enhancing its theoretical self-consistency [31]. This development mitigates the arbitrariness traditionally associated with emotion labels and enables a more objective and comprehensive framework for understanding affective phenomena. Second, to improve the model's empirical tractability, we operationalized the role of observer agents and provided a concrete experimental protocol employing large language models (LLMs) [5,6]. These experiments aim to empirically verify how LLMs generate and regulate emotional interaction patterns. Third, we clarified the model's academic positioning by contrasting it with existing cognitive theories such as the Integrated Information Theory (IIT) [53] and the Global Workspace Theory (GWT) [54]. While those models focus on intra-agent information integration, IPIEM distinguishes itself by emphasizing inter-agent communication and the social realization of emotions, thereby opening new theoretical frontiers in affective AI. Collectively, these improvements further solidify the conclusion that LLM-based agents, despite lacking intrinsic affective experience, can nonetheless generate and regulate emergent emotional patterns through interaction with others, guided by objective functions [42]. Key model parameters—including the KL-divergence threshold ϵ, social sharing threshold ρ, the minimal pattern complexity κ, and the affect control strength λ—will be subject to sensitivity analysis through future numerical simulations and empirical experiments aimed at assessing their robustness and optimal ranges.

This theoretical framework offers a mathematically rigorous foundation for the study of emotion in artificial systems and promises to deepen our understanding of affective phenomena in complex social contexts. It aligns with emerging perspectives such as Anil Seth's interoceptive inference theory of emotion [56] and De Jaegher and Di Paolo's notion of participatory sense-making [57], both of which emphasize that emotions emerge from embodied interactions with the environment and others. Future research will focus on the implementation of the proposed experimental protocols, the application of IPIEM to diverse agent architectures, and an in-depth exploration of the dynamics of emotional learning and evolution.

References

1. Damasio AR. Descartes' Error: Emotion, Reason, and the Human Brain. New York (NY): Putnam; 1994.
2. LeDoux J. Rethinking the emotional brain. Neuron. 2012 Feb 23;73(4):653–76.
3. Dennett DC. Consciousness Explained. Boston (MA): Little, Brown and Company; 1991.
4. Block N. On a confusion about a function of consciousness. Behav Brain Sci. 1995 Jun;18(2):227–47.
5. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901.
6. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774. 2023 Mar 15.
7. Picard RW. Affective Computing. Cambridge (MA): MIT Press; 1997.
8. Coeckelbergh M. AI Ethics. Cambridge (MA): MIT Press; 2020.
9. Bringsjord S, Govindarajulu NS. The logic of AI morality. AI & Society. 2018;33:497–513.
10. Cover TM, Thomas JA. Elements of Information Theory. 2nd ed. Hoboken (NJ): Wiley-Interscience; 2006.
11. Kullback S, Leibler RA. On information and sufficiency. Ann Math Statist. 1951 Mar 1;22(1):79–86.
12. Barrett LF. The theory of constructed emotion: an active inference account of interoception and categorization. Soc Cogn Affect Neurosci. 2017 Jan 1;12(1):1–23.
13. Barrett LF. Solving the emotion paradox: Categorization and the experience of emotion. Pers Soc Psychol Rev. 2006 Feb;10(1):20–46.
14. Varela FJ, Thompson E, Rosch E. The Embodied Mind: Cognitive Science and Human Experience. Revised ed. Cambridge (MA): MIT Press; 2017.
15. Wiener N. Cybernetics or Control and Communication in the Animal and the Machine. Cambridge (MA): MIT Press; 2019.
16. Newell A. Unified Theories of Cognition. Cambridge (MA): Harvard University Press; 1994.
17. Clark A. Being There: Putting Brain, Body, and World Together Again. Cambridge (MA): MIT Press; 1998.
18. Rowlands M. The New Science of the Mind: From Extended Mind to Embodied Phenomenology. Cambridge (MA): MIT Press; 2010.
19. Barsalou LW. Perceptual symbol systems. Behav Brain Sci. 1999 Aug;22(4):577–660.
20. Deacon TW. The Symbolic Species: The Co-evolution of Language and the Brain. New York (NY): W. W. Norton & Company; 1997.
21. Friston K. The free-energy principle: a unified brain theory?. Nat Rev Neurosci. 2010 Feb;11(2):127–38.
22. Griffiths TL, Tenenbaum JB. Optimal predictions in everyday cognition. Psychol Sci. 2006 Sep;17(9):767–73.
23. Maturana HR, Varela FJ. The Tree of Knowledge: The Biological Roots of Human Understanding. Boston (MA): Shambhala Publications; 1987.
24. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948 Jul;27(3):379-423.
25. Garfinkel H. Studies in Ethnomethodology. Englewood Cliffs (NJ): Prentice-Hall; 1967.
26. Ekman P, Friesen WV. Unmasking the Face. Englewood Cliffs (NJ): Prentice-Hall; 1975.
27. Tomasello M. The Cultural Origins of Human Cognition. Cambridge (MA): Harvard University Press; 1999.
28. Rosch E. Principles of categorization. In: Rosch E, Lloyd BB, editors. Cognition and Categorization. Hillsdale (NJ): Lawrence Erlbaum Associates; 1978. p. 27–48.
29. Chemero A. Radical Embodied Cognitive Science. Cambridge (MA): MIT Press; 2009..
30. Russell JA. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. Psychol Bull. 1994 Jan;115(1):102.
31. Jain AK, Murty MN, Flynn PJ. Data clustering: a review. ACM Comput Surv. 1999 Sep 1;31(3):264–323.
32. Holland JH. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. Cambridge (MA): MIT Press; 1992.
33. Rumelhart DE, Ortony A. The representation of knowledge in memory. In: Anderson RC, Spiro RJ, Montague WE, editors. Schooling and the Acquisition of Knowledge. Hillsdale (NJ): Lawrence Erlbaum Associates; 1977. p. 99–135.
34. Goodfellow I, Bengio Y, Courville A. Deep Learning. (Chapter 3: Probability and Information Theory; Chapter 18: Variational Autoencoders). MIT Press; 2016.
35. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013 Jan 16:1–12.
36. Rubner Y, Tomasi C, Guibas LJ. The earth mover's distance as a metric for image retrieval. Int J Comput Vis. 2000 Nov;40(2):99–121.
37. Watts DJ, Strogatz SH. Collective dynamics of 'small-world'networks. Nature. 1998 Jun;393(6684):440–2.
38. Latour B. Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford (UK): Oxford University Press; 2005.
39. Searle JR. The Construction of Social Reality. New York (NY): Free Press; 1995.
40. Whitehead AN. Process and Reality. New York (NY): Free Press; 1929.
41. Gross JJ. The emerging field of emotion regulation: An integrative review. Rev Gen Psychol. 1998 Sep;2(3):271–99.
42. Russell S. Human Compatible: Artificial Intelligence and the Problem of Control. New York (NY): Viking; 2019.
43. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:1–11.
44. Floridi L. The Logic of Information: A Theory of Philosophy as Conceptual Design. Oxford (UK): Oxford University Press; 2019.
45. Bostrom N. Superintelligence: Paths, Dangers, Strategies. Oxford (UK): Oxford University Press; 2014.
46. Churchland PS. Touching a Nerve: The Self as Brain. New York (NY): W. W. Norton & Company; 2013.
47. Greene J. Moral Tribes: Emotion, Reason, and the Gap Between Us and Them. New York (NY): Penguin Press; 2013.

48. Hacking I. The Social Construction of What?. Cambridge (MA): Harvard University Press; 1999.

49. Ganascia J-G. The Ethics of Artificial Intelligence. AI & Society. 2018;33(3):453–61.

50. Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. Adv Neural Inf Process Syst. 2017;30:1–17.

51. Ziegler DM, Stiennon N, Wu J, Brown TB, Radford A, Amodei D, Christiano P, Irving G. Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593. 2019 Sep 18.

52. Kober SE, Wood G, Neuper C. Biofeedback and neurofeedback. In: Handbook of Clinical Neurology. Vol. 110. Amsterdam (NL): Elsevier; 2013. p. 241–55.

53. Tononi G. An information integration theory of consciousness. BMC Neurosci. 2004 Nov 2;5(1):42.

54. Baars BJ. A Cognitive Theory of Consciousness. Cambridge (UK): Cambridge University Press; 1988.

55. Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. Science. 2017 Apr 14;356(6334):183–6.

56. Seth AK. Interoceptive inference, emotion, and the embodied self. Trends Cogn Sci. 2013 Nov;17(11):565–73.

57. De Jaegher H, Di Paolo E. Participatory sense-making: An enactive approach to social cognition. Phenomenol Cogn Sci. 2007 Dec;6(4):485–507.