The Emotional Foundation of Human-AI Interaction: Theoretical Insights from Evolutionary Continuity and Interspecies Emotional Communication (Postprint)

Authors: Liu Chongyi, Yin Bin, Yin Bin

Date: 2025-05-10T14:22:22+00:00

Abstract

The era of Artificial General Intelligence (AGI) is imminent, compelling us to reassess human-AI interaction, particularly through emotional communication. This study synthesizes insights from evolutionary biology, comparative psychology, and artificial intelligence development, advocating a paradigm shift beyond traditional human-like cognitive processes. The research emphasizes the universality of emotional pathways, which is manifested across different species. We introduce three models of emotional interaction—the emotional threshold model, the dynamic set-point model, and the emotional schema model—which are derived from in-depth analysis of interspecies emotional interaction phenomena and potential mechanisms. These models provide a roadmap for designing AI interfaces that align with human emotional experiences, elucidating pathways for establishing trust, intuition, and mutual recognition between machines and humans. By further clarifying the concept of the "Large Emotional Model," we envision a future where artificial intelligence can not only interpret but also comprehend the emotions of human partners, paving the way for a revolutionary collaborative paradigm between artificial intelligence and humans.

Full Text

Affective Foundations in AI-Human Interactions: Theoretical Insights from Evolutionary Continuity and Interspecies Affective Communication

Chongyi Liu, Bin Yin†
School of Psychology, Fujian Normal University, Fuzhou, Fujian 350117

Abstract

The imminent arrival of Artificial General Intelligence (AGI) compels us to reevaluate artificial intelligence's interactions with humans, particularly through the lens of affective communication. This study synthesizes insights from evolutionary biology, comparative psychology, and AI development, advocating for a paradigm shift beyond traditional anthropocentric cognitive processes. We emphasize the universality of affective pathways, evident across diverse species. We introduce three affective interaction models—the Affective Threshold Model, Dynamic Set-Point Model, and Affective Schema Model—derived from in-depth analysis of interspecies affective interaction phenomena and their underlying mechanisms. These models provide a roadmap for designing AI interfaces that align with human affective experiences, elucidating pathways for establishing trust, intuition, and mutual recognition between machines and humans. By further clarifying the concept of a "Large Affective Model," we envision a future where AI can not only interpret but also understand the emotions of human partners, paving the way for a revolutionary collaborative paradigm between AI and humans.

Keywords: affective communication; AI-human interaction; affective threshold model; dynamic set-point model; affective schema model; large affective model

The English version of this article is cited as: Liu, C. Y., & Yin, B. (2024). Affective foundations in AI-human interactions: Insights from evolutionary continuity and interspecies communications. Computers in Human Behavior, 161, 108406. doi: 10.1016/j.chb.2024.108406
†Corresponding author. Email: byin@fjnu.edu.cn

1. Introduction

The rise of the AI era, marked by the emergence of large language models (LLMs), represents a significant leap in artificial intelligence. Models like OpenAI's GPT-4 have demonstrated capabilities rivaling or exceeding human expertise across diverse domains including drug discovery, biology, mathematics, and chemistry (AI4Science & Quantum, 2023). Notably, GPT-4 exhibits human-like emotion regulation abilities, suggesting AI is making important strides toward adjusting responses based on emotional cues (Zhao et al., 2023). These advances signal AI's technological progression toward the Artificial General Intelligence (AGI) era, with characteristics of early AGI becoming increasingly evident (Morris et al., 2023).

Numerous reviews defining and classifying AGI emphasize systems designed to match or surpass human capabilities in generalization and performance (e.g., Aher et al., 2023; Aminah et al., 2023; Broekens et al., 2023; Bubeck et al., 2023; Dwivedi et al., 2023; Li et al., 2023; Rao et al., 2023; Schueller & Morris, 2023; Yongsatianchot et al., 2023). To achieve these ambitious goals, AI has long been intertwined with neuroscience and psychology, drawing crucial technical inspiration from these disciplines (Hassabis et al., 2017). Over the past decade,

the resurgence of neural networks (i.e., deep learning) has driven revolutionary advances in AI (LeCun et al., 2015; Schmidhuber, 2015). As the term "neural network" suggests, these models are inspired by biological nervous systems, simulating connections and information transmission between neurons (Hassabis et al., 2017). Deep learning algorithms have also benefited from insights in cognitive psychology. For instance, attention mechanisms—which monitor and filter inputs at each computational step to improve target classification accuracy in noisy environments—have become integral components of AI architectures (Niu et al., 2021). Furthermore, recent developments in meta-learning frameworks enable AI models to "learn to learn," adapting from past experiences to better address new challenges (Binz et al., 2023). This concept originated from animal studies (Harlow, 1949) and was further explored in developmental psychology (Adolph, 2005; Kemp et al., 2010), highlighting the profound impact of interdisciplinary research on AI development.

However, while these advances primarily enhance AI's cognitive capabilities, they often overlook a critical aspect: the seamless integration of AI-human interactions (AHI). For AGI to achieve true generality, the dynamic interplay between humans and AI systems must be considered. This requires reevaluating the AHI paradigm to ensure technological progress remains human-centered and constructively integrated into society.

Current trends in AHI involve designing AI systems that mimic human behavior, often leading to anthropomorphism in AI development (Broadbent et al., 2013; Epley et al., 2008; Li & Sung, 2021; Salles et al., 2020). This approach is based on trait attribution theory, which posits that people tend to ascribe human characteristics to non-human entities (Epley et al., 2007; Ruijten et al., 2019; Waytz et al., 2010). However, this theory oversimplifies individual uniqueness and complexity, proving inadequate for explaining the nuances of AHI, particularly in scenarios requiring sustained real-time interaction (Clark & Fischer, 2023). Contrasting with traditional views, Clark and Fischer (2023) propose a novel interpretation of AHI, suggesting that individuals do not perceive social robots as autonomous agents per se, but rather as depictions of agents. This perspective explains why people can develop genuine affective responses toward social robots deemed inanimate—these responses stem from personal interpretation and the imagination imbued in the robot's behavior. Social robots equipped with advanced affective systems can interact with humans in more natural, personalized ways. By effectively mimicking social behaviors and responses, they enhance their role as social entities in domains like healthcare and education. Nevertheless, the complexity of real-world scenarios demands that affective robots not only recognize but also understand users' affective cues based on comprehensive behavioral and contextual analysis (Ong et al., 2021). Yet past efforts to replicate human cognitive functions have failed to satisfactorily address the challenges of affective computing. Current affective computing models excel at recognizing patterns in emotional expression but do not truly understand underlying emotions or their causal contexts (Ong et al., 2021). For example, while facial emotion recognition technologies like Deep Region and

Multi-label Learning (DRML) can accurately analyze emotional expressions (K. Zhao et al., 2016, pp. 3391-3399), they often lack the flexibility needed to adapt to different contexts—flexibility essential for genuine emotional understanding.

To facilitate more natural, flexible, and effective AI-human interactions, a robust theoretical framework is essential. We advocate developing AI systems that can be considered "comprehensive affective entities," not merely based on cognitive and linguistic capacities. This paradigm shift involves embracing broader affective processes observed in biological organisms—organisms that display various cognitive capabilities while engaging in rich affective interactions. This approach aims to identify commonalities across different forms of biological intelligence, much like humans recognize consciousness in other species while acknowledging their unique animal characteristics. Research by Merkies et al. (2022) supports this deeper understanding of biological intelligence, demonstrating that humans can comprehend emotional signals across different species.

Fundamentally, affect is experienced from a first-person perspective (Schiller et al., 2024). Therefore, our scope extends beyond humans to include all sentient beings capable of experiencing affect. To systematically explore these universal affective pathways, we developed three core models: the Affective Threshold Model, Dynamic Set-Point Model, and Affective Schema Model. These models aim to investigate the evolutionary development of affect and its manifestation across species with varying cognitive levels, providing insights into complex general affective states.

When constructing affective computing systems, strict biological accuracy is not our primary goal; practicality and functionality guide our choices. However, understanding the fundamental principles of biological affective systems at computational and algorithmic levels (Marr & Poggio, 1976) provides valuable guidance for our work. By integrating these biological insights into AGI development, we aim to create AI systems capable of genuine emotional resonance. This development strategy seeks to enhance AI-human interactions by fostering empathy, trust, and intuitive connection, fundamentally transforming the interaction landscape.

The remainder of this paper is organized as follows: Section 2 discusses the evolutionary continuity of affective pathways, laying the foundation for advanced AI-human interaction models. Section 3 introduces three potential models for cross-species affective interaction. Section 4 applies these models to AI-human interface design, proposing the "Large Affective Model," while Section 5 explores the significance of affective understanding in advancing AI alignment. The paper concludes in Section 6 with reflections on connecting insights from biological evolution to future AI development.

2. Evolutionary Continuity of Affective Pathways: Laying the Foundation for Advanced AI-Human Interaction Models

Before proceeding, it is necessary to clarify fundamental concepts such as the definition of "affect" and its distinction from "cognition." Historically, scholars across disciplines have held different assumptions about affective phenomena (Izard, 2009; Lazarus, 1984; Zajonc, 1984), leading to significant variations in phenomenological analysis, measurement methods, and experimental design. In an important integrative effort, Schiller et al. (2024) synthesized these diverse perspectives into a unified set of teleological principles. They define "affect" as the mechanism by which neurobiological organisms demonstrate their relevance in the world by engaging in survival-oriented activities. This concept posits that all affective phenomena ultimately serve to ensure survival capacity and influence subsequent changes in affective characteristics based on an organism's affective concerns. Within this framework, abstract cognitive activities such as language, thought, and reasoning are viewed as components of the affective landscape, serving adaptive functions and facilitating environmental engagement. We apply Schiller et al.'s (2024) comprehensive understanding of affect in our discussion to broadly explore cross-species affective interactions.

Furthermore, it is necessary to clarify our use of the term "affect" rather than "emotion" to avoid confusion when both terms appear in the text. Affect encompasses broad psychological states including feelings, emotions, moods, preferences, desires, fears, joy, and other subtle sentiments. It represents the fundamental, instinctive reactions an entity experiences—reactions that are primitive and typically 未经 cognitive evaluation. In contrast, emotion is a subset of affect, usually more structured and complex, influenced by cultural and contextual factors, and involving higher-level cognitive processing and interpretation. In our discussion, we primarily use the term "affect" to emphasize these broad, primitive psychological experiences that have been preserved throughout biological evolution and are crucial to biological entities' behavior. This comprehensive focus on affect allows us to deeply explore the universality of these experiences across species, providing critical insights into their evolutionary continuity and implications for developing advanced AI-human interaction models.

2.1 Affect-Driven Biological Agents: Evolution, Goal-Directed Behavior, and Complex Affective Networks

In both natural systems and artificial constructs like robotic devices and computational systems, all entities can be broadly categorized as information processing systems. Wang et al. (2022) define intelligence as the capacity of such systems to operate and adapt in environments with limited knowledge and resources. This definition does not require agents to always succeed but emphasizes their proficiency in optimizing the use of available knowledge and resources within existing constraints. Therefore, measuring an information processing sys-

tem's intelligence should be based on its ability to handle ill-defined problems rather than well-defined ones. In nature, animals are typical representatives of biological agents that exhibit not only passive reactive tendencies to environmental stimuli but also active, goal-directed behavior. They utilize and explore their environment, overcoming obstacles to achieve their goals. The affordances of objects in the environment are crucial, as biological agents must skillfully exploit these affordances to rapidly solve problems aligned with their goals and needs (Roli et al., 2022). For instance, some primates like chimpanzees and gorillas use branches as tools to obtain insects or stones as hammers; foxes skillfully use terrain to approach and capture prey; even ants collaborate to use environmental materials to build complex nests.

The essence of this environmental "creativity" is driven by goal-directed motivation, which differs from mathematical, probabilistic, geometric, or logic-oriented trends that lack physical considerations and are inconsistent with goal-directed paradigms (McShea, 2023). In biological agents, this goal-directed mechanism is intimately linked to their innate affective systems. Affective science explores these neurobiologically-based subjective phenomena, including both implicit and explicit phenomena, striving to clarify their significance for sentient beings. This inquiry follows teleological principles that ask "why" affective phenomena occur, suggesting these behaviors' ultimate purpose is ensuring survival capacity (Schiller et al., 2024). This concept implies that biological systems' internal needs are maintained within specific homeostatic ranges to support life (Damasio & Carvalho, 2013). The fundamental goals underlying all behavior are maintaining stable states and propagating biological information through reproduction. Affect encompasses a broad spectrum, from simple interests and moods to complex drives, desires, ethical values, and aesthetic feelings (Panksepp, 2012). Cross-species research demonstrates that core affective systems exhibit continuity across biological evolution, emphasizing affect's role as a universal "currency" in decision-making across functional domains: if current behavior generates pleasure, continue; if discomfort is felt, switch to escape and avoidance modes (Mendl & Paul, 2020).

It must be recognized that the goal-directed nature of biological agents inherently includes the possibility of failure. Affect-guided actions do not always align with the organism's survival or reproductive interests. While choice does not equal correctness, the ability to make "choices" based on internal organization is a fundamental aspect of intelligence. Biological systems' intentionality stems from their autonomy, enabling organisms to maintain and create conditions favorable for survival through their own activities (Mossio & Bich, 2017). This autonomy can be observed even in single-celled organisms like amoebas, which exhibit basic discriminative abilities and respond to environmental stimuli to achieve survival and reproduction.

In biological evolution, organisms gradually developed temporal delay mechanisms between sensation and response, marked by the emergence of brains and the capacity to construct internal representations of external stimuli within the

nervous system. This mechanism gave rise to planning-based decision-making grounded in cognitive models, enabling organisms to formulate prospective action plans. Such rationalized decisions often generate internal conflicts with affective drives, reflecting deep evolutionary tensions within the nervous system. The mid-20th century cognitive revolution in psychology largely equated cognition with the brain's computational functions (Cromwell & Panksepp, 2011). However, the importance of affect in biological agents cannot be underestimated. As Panksepp (2012) proposed, mammals' subcortical regions are closely associated with the formation of affective networks. These networks consist of basic affective responses to unconditional stimuli combined with higher-level affective components involving more complex cognitive processes. In summary, rationality, logic, computation, and other non-affective capacities necessarily serve affective capacities (McShea, 2017). Cognitive and volitional systems are interdependent, constituting a coherent agent. Through evolutionary processes, this integrated system has evolved diverse affects, constructing a broad and complex affective network. Animal brains evolved not only for representation and logical computation but more critically for maintaining internal environmental stability and organismic function, ensuring life's growth, survival, and reproduction. This shared goal establishes a universal trend in the biological world: the fusion of affective, cognitive, and volitional systems into a synergistic entity driving life's continuous reproduction and evolution. The dominance of affective systems in animal brains reflects not only their autonomy but also patterns of information exchange and interaction between individuals. Next, we explore how this shared behavioral logic influences interactions across different life forms.

2.2 Universality of Affective Interaction: Survival Instincts in Biological Entities

Affect plays a crucial role in interactions between biological entities, both within human society and the broader natural world. These affects are foundational for triggering intentional and unintentional behaviors. In cases of intentional behavior, affective states can serve as catalysts that directly influence actions without relying on beliefs and desires (Ong et al., 2016). Consequently, the ability to infer a entity's probable affective states, intentions, and desires from its behavioral expressions, and to predict subsequent actions based on these inferred states, constitutes a critical foundation for interactions between biological entities. For example, in daily interactions, we infer others' emotions by interpreting their facial expressions and body language to predict their subsequent actions. This represents a form of inverse modeling where we attribute intentional behavior to an agent's beliefs and desires (Baker et al., 2009). When evaluating others' affective states, observers engage in a "third-person evaluation" process, assessing from the agent's perspective whether outcomes align with the agent's goals and expectations (Ong et al., 2019). However, this approach may be inaccurate, as it relies on the observer's interpretation of the agent's prior beliefs and desires, which may not accurately reflect the agent's

true mental state.

Efforts to understand others' affective states are considered an aspect of cognition, often termed affective cognition (Ong et al., 2015). Many studies have developed computational models for inferring others' affective states (e.g., Baker et al., 2017; Houlihan et al., 2023; Ong et al., 2019; Saxe & Houlihan, 2017; Wu et al., 2018), aiming to achieve precise affective attribution. In these models, affect is viewed as a personal interpretation of external stimuli, with additional contextual information improving attribution accuracy. Observers' ability to distinguish between facial expression features and structural features improves with dynamic changes. When facial expression changes correspond to the temporal structure of external events, observers can more accurately infer the emotional dimensions of events (Saxe & Houlihan, 2017). However, this complex, abstract, and highly formalized cognitive computation fails to capture the universality and essence of biological affective interaction. In the animal kingdom, recognizing others' affective intentions is a crucial survival skill, manifesting as a behavior that transcends logical reasoning. An anecdote from Bertrand Russell's *The Problems of Philosophy* highlights the limitations of computational intelligence in survival contexts. The story tells of a farm turkey that initially remained vigilant but grew increasingly complacent as it misinterpreted the farmer's daily feeding behavior through Bayesian probability analysis as a positive affective attitude. However, this so-called intelligence did not prevent its slaughter on Christmas Eve. This example emphasizes that while computational and logical reasoning skills are valuable, they should ultimately serve thoughts, preferences, and visions—the true drivers of environmental interaction (McShea, 2017). The application of such advanced cognitive abilities, often considered hallmarks of human intelligence, may not align with the evolutionary reality and survival needs of biological organisms. The turkey story profoundly reminds us that in the biological world, survival often depends on simple emotion-based decisions rather than complex computational analysis.

Therefore, simpler, more universal affect-action mechanisms likely exist, enabling biological entities to rapidly identify affective cues and respond quickly during interactions. While these mechanisms may be imprecise due to information loss, they possess sufficient universality and breadth to provide a foundation for communication across different species.

2.3 Exploring Communicative Intent Beyond Cognitive Attribution: Insights from Affective Phenomena in Cross-Species Interactions

To facilitate meaningful communication between individuals, a shared mental framework is essential, functioning similarly to an information decoder that enables parties to understand each other's intentions and beliefs (O'Madagain & Tomasello, 2022; Scott-Phillips, 2015). This complex meta-psychological structure is considered a fundamental prerequisite for language evolution and the core of effective communication, encompassing speakers' intentions to convey information, expectations that listeners understand these intentions, and corre-

sponding responses. However, this theory—primarily based on human linguistic communication—lacks universality when extended to broader species, facing three main challenges: first, it demands excessive cognitive resources, requiring real-time mental state attribution from both signal senders and receivers—a complex ability even difficult for adults, let alone young children and other primates (Moore, 2016). Second, it underestimates the importance of non-verbal communication, which in both natural and human societies often relies on actions and behaviors. For instance, non-primate animals consider recipients' attentional states, gesturing only when recipients are appropriately attentive (Leavens et al., 2005). Third, the theory overemphasizes mental attribution in communication, neglecting situations where understanding does not require reasoning about speakers' intentions (Townsend et al., 2017). These insights highlight the limitations of traditional communication theories when applied across species, emphasizing the need for a more inclusive approach.

Townsend et al. (2017) propose a new framework for communicative intent that avoids reliance on mental state attribution. This framework comprises three conditions: first, goal-directed behavior, where communicators have clear goals and understand the relationship between actions and outcomes; second, voluntariness, where communicators intentionally signal to achieve specific goals; and third, communicative behavior that can repeatedly and consistently point to something in the world aligned with the communicator's intentions (Slaby & Stephan, 2008). For example, studies on rodents show that free-moving rats persistently attempt to rescue trapped companions (Bartal et al., 2011, 2016), indicating they understand trapped rats' affective states and the reasons behind these affects—aligning with Townsend's communication framework. Unlike human-specific cognitive intentions, affective intentionality universally exists across species and can be transmitted interspecifically to some degree. From an evolutionary perspective, affective representations essentially map organisms' primitive intentions to seek benefits and avoid harm. The perception of and feedback to affective intentions constitute a universal communication paradigm across species, achieving efficient interaction through information simplification and dimensional compression.

Cross-species cooperative behavior demonstrates how this affect-based communication model supports interactions between biological agents and their environment. Cooperation within species, such as division of labor in meerkat groups or collaborative hunting in dolphins, often receives attention for its cognitive complexity. For instance, Łopuch and Popik (2011) and other scholars (Hauser et al., 2009; Nowak, 2012) argue that cross-species cooperation requires high cognitive ability. Conversely, Brown (2006) proposes that intense local competition drives trait differentiation among cooperators, evolving local group interactions into robust cross-species symbiotic relationships. This differentiation reduces cheaters' success rates and improves resource utilization efficiency. For example, cooperative hunting between North American coyotes and badgers demonstrates this principle: each species contributes unique skills, thereby improving hunting efficiency and success rates in challenging environments (Minta

et al., 1992). In such systems, associative learning and recognition of affective intentions play key roles in establishing and maintaining cross-species relationships. Both species must recognize each other's behavior as positive affective signals and associate these behaviors with previous hunting successes.

In summary, the construction of affective interaction models among biological entities highlights the theoretical boundaries of cognitive attribution paradigms in explaining cross-species interactions, with limitations particularly significant along dimensions of species difference. As a meta-linguistic system for interspecies communication, affective representation mechanisms achieve primitive information exchange paradigms beyond cognitive frameworks through information dimensionality reduction and pattern compression strategies. Moreover, cross-species cooperative behavior demonstrates that affect-based communication is integrated into the logic of biological interactions, challenging traditional views that emphasize cognitive complexity in such interactions.

3. Three Potential Models for Cross-Species Affective Interaction

In the preceding sections, we explored the universality of affect among biological agents and the fundamental principles of affective communication, characterized by universally present action-activation systems across life forms. Empirical research shows that interactions within and between species are rapid and direct, where one organism's affective state immediately triggers corresponding behaviors in others. This reflects a bottom-up activation process that does not rely on complex meta-cognitive strategies to infer intentions. From an evolutionary perspective, such rapid response systems provide survival advantages, enabling organisms to quickly defend against environmental threats. For example, responding to others' fear or distress signs by activating one's own defense mechanisms allows individuals to prepare for potential dangers, thereby avoiding direct encounters with potentially lethal situations (Keysers et al., 2022). Based on these insights, we propose three models describing affective interaction derived from phenomenological studies of biological behavior: the Affective Threshold Model (ATM) (Figure 1 [Figure 1: see original paper]), Dynamic Set-Point Model (DSPM) (Figure 2 [Figure 2: see original paper]), and Affective Schema Model (ASM) (Figure 3 [Figure 3: see original paper]). Each model includes detailed theoretical background, explanatory contexts, and validation methods. Together, these models aim to deepen our understanding of cross-species affective interactions and provide an effective framework for advancing affective modeling. We believe this will offer important perspectives and a crucial foundation for addressing the complexities and challenges of AI-human interactions.

3.1 Affective Threshold Model (ATM)

The Affective Threshold Model elucidates the mechanisms of positive affective interactions between agents with different characteristics, ecological niches, and cognitive abilities. The model's core premise is that environmental stimuli and affective cues from interactants simultaneously trigger positive-valence experiences in both parties, ensuring pleasure generation in specific interaction contexts. Research shows that individuals vary in their sensitivity to affective stimuli (e.g., Hyett et al., 2014; Karmon-Presser et al., 2018; Sokolov & Boucsein, 2000).

Signal Detection Theory (SDT) provides a powerful analytical framework for sensory research, describing perception formation by assuming stimuli generate internal sensations that must exceed certain thresholds to be perceived (Mamassian, 2016). SDT comprises two experimental paradigms: detection paradigms where participants discriminate weak stimuli, measuring sensitivity (d'); and discrimination paradigms where participants distinguish between two similar stimuli, with d' representing the incremental sensation produced by stronger versus weaker stimuli (Wixted, 2020). This theory also applies to affect generation processes, treating sensitivity and judgment criteria as latent variables inferred from observed hit and false alarm rates (Chang et al., 2015; Nielsen & Kaszniak, 2006). The SDT model of affective triggering posits that affect generation involves inherent uncertainty, influenced by mood and contextual factors, thus requiring the ability to discriminate between different intensity affective stimuli and establishing reporting thresholds (Karmon-Presser et al., 2018).

Comparing affective sensation research with sensory processes, both share the core view that information transmission requires integration and must exceed certain thresholds to trigger higher-level responses. Neurophysiological research confirms this view. Huzard et al.'s (2022) study found that activating specific somatosensory neurons (C-LTMRs) in mice directly influences their emotional responses and prosocial behavior. Modulating C-LTMRs' excitability can simulate different tactile experiences, thereby affecting mice's responses to social stimuli. This finding emphasizes the importance of stimulus accumulation reaching levels that activate neurons for affective experience and social behavior, where single neuron action potential generation and information integration across multiple neurons constitute the basis of complex affect.

Extending from the SDT model, low-dimensional affective structures—such as subcortical regions related to animal affective behavior and the autonomic and somatic components of human emotion—are more likely to follow the Affective Threshold Model. In contrast, subjectively experienced, complex, cognitively constructed emotions primarily manifest in cortical regions (Sokolov & Boucsein, 2000). Although complex emotions differ between animals and humans, all affective experiences are intrinsically related to physiological sensory signals, with positive and negative evaluations related to survival needs closely linked to approach and avoidance behaviors. This suggests a reliable mapping between

behavior, sensation, and affective evaluation.

For survival, a critical aspect of affective judgment is the perception of safety. Notably, "perceived safety" and "noticed danger" are not antonyms; the opposite of perceived safety is "failure to perceive safety." Counterintuitively, organisms' stress responses are often triggered not by explicit threats but by the absence of safety (Brosschot et al., 2016). This reflects a default threat-perception state, where stress responses are inhibited only when safety is explicitly perceived. Environmental uncertainty maintains the default stress response in an activated state. Essentially, the ability to establish safety in unfamiliar environments represents a tolerance for uncertainty, which may manifest as a tendency to categorize novel stimuli as threats (Carleton, 2012). Figure 1 illustrates the concept of accumulating positive stimuli to reach safety perception thresholds. Figure 1A depicts the psychophysical relationship between positive stimuli in the environment and perceived safety. Empirical research by Berkovich and Meiran (2023) shows that positive affect perception follows Weber's law in centralized information perception, meaning stimulus encoding accuracy decreases with intensity increases. Weber's law has been confirmed across nearly all sensory domains (Akre & Johnsen, 2014; Simen et al., 2016), suggesting that organisms' perception of environmental safety may also follow this principle, with different activation thresholds and response curves reflecting individual uncertainty tolerance. During interactions, organisms with lower safety thresholds can rapidly establish safety and emit positive signals through behavior or emotional contagion, helping organisms with higher safety thresholds convert uncertain environmental information into explicit positive stimuli, ultimately reaching their safety thresholds. Figure 1B illustrates this mutual safety establishment process.

In summary, the Affective Threshold Model provides insights into how organisms generate positive affective experiences during interactions. Through analogy with sensory processes, the model emphasizes the importance of information transmission and perception in affective interactions. Moreover, it highlights the role of safety perception, proposing that—like perception of positive environmental stimuli—safety perception follows Weber's law, with different individuals' thresholds and response curves reflecting their uncertainty tolerance. The model suggests that organisms with lower safety thresholds more easily establish safety, facilitating positive affect propagation and promoting positive interactions with organisms having higher safety perception thresholds.

[Figure 1: see original paper]

3.2 Dynamic Set-Point Model (DSPM)

The Dynamic Set-Point Model elaborates biological agents' functions based on innate value and affective systems. As described in Section 2.1, the primary goals of maintaining comfortable states and ensuring biological reproduction constitute the cornerstone of all affects, shaping the rich and diverse affective

spectrum throughout biological evolution. This spectrum ranges from basic interests and moods to complex drives and desires, with cognitive and volitional systems interdependent and jointly serving these affective goals. Animal brain evolution focuses not merely on representation, reasoning, and computation, but more fundamentally on maintaining internal environmental stability and ensuring organismic growth, survival, and reproduction.

In dynamic interactions with the environment, all animals—from blue whales to ants—distinguish "self" from surroundings through bodily boundaries. Animals that have evolved brain structures possess central nervous systems that interact with the environment through the body and its sensory organs (Kahl & Kopp, 2023). Survival and adaptation require not only cognitive processing of environmental information but also action to generate material causal forces addressing real-world challenges. A critical aspect of this interaction involves interoception—the perception of internal signals from the autonomic nervous system, hormones, immune responses, and organs (Craig, 2003). Damasio and Carvalho (2013) further developed this concept, distinguishing between feelings and emotions. The former are psychological experiences accompanying bodily states. This body-based sensation is crucial for decision-making: we predict possible outcomes of our actions, and the feelings these outcomes generate form the basis of our actions.

Mismatch between subjective interoceptive sensitivity and objective accuracy generates error signals when predicting internal bodily information. Feelings are not merely representations of visceral sensory information; they involve a bidirectional pathway. The brain makes top-down predictions about bodily states that interact with bottom-up interoceptive neural signals, enabling error correction (Critchley & Garfinkel, 2017). For example, the brain can instruct bodily movements, but during environmental interaction, how this behavior will develop and whether it will encounter obstacles cannot be fully predicted in advance. Therefore, optimally, neural signals reaching specific muscles also contain errors that propagate upward during movement for timely adjustment. In short, the interaction between brain and body is central to this model. In a continuous feedback loop, body and brain exchange signals; this loop involves the autonomic nervous system, which responds to external inputs and internal sensory states, enabling our various arousal states and "fight-or-flight" responses to function or adjust. It is through this mechanism that the autonomic nervous system influences homeostatic regulation. Allostasis—the predictive regulation mechanism of homeostasis—explains physiological and behavioral regulation processes (Schulkin & Sterling, 2019). Imagine being in a hot environment: if we cannot predict environmental temperature's impact on body temperature and adjust accordingly (e.g., quickly leaving the environment), we might die. Allostasis is a prospective regulation mechanism where the body actively prepares for disturbances that have not yet occurred. We feel hot before heatstroke, thirsty before dehydration, hungry before fainting. It also explains stress state generation: when facing unexpected environmental stimuli, these evolved autonomous responses overwork, thereby harming the body. Nevertheless, these complex

interoceptive processes continuously operate to maintain physiological regulation and basic survival, typically unconsciously. Figure 2A extends this theory, further proposing the Dynamic Set-Point Model based on maintaining homeostatic goals to better illustrate how biological agents induce bodily changes and drive behavior through value systems manifested as affect. We argue that to adapt to environmental changes, homeostatic goals should be a set of dynamically adjustable data structures, encompassing physical or biochemical states and psychological or affective states that organisms need to maintain. In different situations, the weights of these needs vary. The brain predicts and evaluates current bodily input information, generating corrective value signals that act on motor control systems and the autonomic nervous system, achieving negative feedback regulation.

The brain's interoceptive predictive processing follows Bayesian active inference principles. Bayesian reasoning updates prior beliefs or hypotheses based on probabilistic reasoning (Knill & Pouget, 2004). The brain uses interoceptive information to interpret stimuli, regulating allostasis to minimize prediction errors about the environment. This continuous prediction-correction process from past to future experiences shapes our current perception of internal bodily sensations (Barrett & Simmons, 2015). Interactions between different biological agents are essentially dynamic information exchange processes within closed systems, where the interaction itself constitutes the recursive construction of system closure. Notably, this closure is never a static, 固化 boundary but continuously reconstructs through the historical unfolding of self-organizing networks: the hierarchical structure currently emerging in the system is essentially the product of iterative evolution of past organizational patterns (Roli et al., 2022). Figures 2B, C, and D reflect the adjustment and change of internal setpoints during affective interactions between two agents. Through environmental changes and introduction of new stimuli, the weight of needs in the monitoring model changes, enabling both parties to gradually explore higher-level social interaction needs.

This model provides a theoretical framework for explaining how biological agents adjust internal stable states based on affective value systems to adapt to environmental changes. It emphasizes the critical role of affective systems in regulating biological agents' internal stability. Through affective value systems, biological agents can dynamically adjust internal stable states according to environmental changes, enabling better adaptation to different situations and needs. When biological agents face constantly changing external environments, they require adaptability to maintain internal stability and make effective behavioral choices. The Dynamic Set-Point Model provides biological agents with an affect-driven mechanism for internal adjustment based on environmental changes, thereby increasing their advantages for survival and reproduction in complex environments.

[Figure 2: see original paper]

3.3 Affective Schema Model

Our third affective interaction model—the Affective Schema Model—delves deeper into representing learned affective response patterns and emphasizes neural plasticity's role as the foundation of associative learning. This learning mode, including classical and operant conditioning, is essential for survival across numerous species, including those without central brains. Bielecki et al. (2023) demonstrated that even cnidarians like jellyfish with distributed nervous systems can perform operant conditioning. Their research trained the box jellyfish *Tripedalia cystophora* to associate low-contrast objects with collision risk, challenging traditional notions that operant conditioning requires conventional central nervous systems and suggesting that complex neural processes like operant conditioning may be fundamental features of neural circuits, even in simpler life systems. Conditioning and stimulus generalization lead to individual differences in affective responses, helping organisms approach potential rewards and avoid threats. Neurophysiological explanations of stimulus generalization involve establishing contextual memories that gradually reduce hippocampal dependence over time, with the cerebral cortex independently representing commonalities of past events (Heller, 2020; Morrissey et al., 2017). The prefrontal cortex uses past memories for prediction and influences perceptual and motor systems (Miller & Cohen, 2001), echoing the top-down predictive processes mentioned in the Dynamic Set-Point Model. Over time, memories fade, and stimulus and background features triggering affective responses become blurred, with some structural frameworks and salient features becoming generic representations of positive or negative stimuli in the environment. For example, in Pavlov's classic experiments, if dogs were only exposed to specific bell frequencies during training, they might also show conditioned responses when hearing completely different but similar frequencies.

Heller (2020) treats the abstraction process of conditioned stimuli as schema formation. For different but overlapping events, common nodes can be extracted to define a schema's non-specific core experience. This differs from Piaget's constructivist epistemology, where schemas form through individuals' interactions with reality, encompassing knowledge about self, others, and the world, representing all life experiences accumulated during world interaction (Y.-X. Wang & Yin, 2023; Yin et al., 2022). The latter clearly integrates all experiences at a higher dimension rather than generalizing specific conditioned stimuli. However, the two are not conflicting; schemas are hierarchical structures where each higher-level schema consists of sub-schemas (Bein & Niv, 2023). When processing real-world tasks, this hierarchy can decompose overall goals into different subtasks, achieving maximum efficiency through cross-task sharing. When encountering new stimuli or situations, organisms infer entire events based on partial cues and complete other elements, a process called pattern completion (Ngo et al., 2021). Unlike the Dynamic Set-Point Model, the Affective Schema Model triggers organisms' approach/avoidance behaviors by selectively activating different affective schemas based on information obtained from homeostatic goals

(Figure 3A). Environmental feedback information confirms or falsifies schema-based predictions, with deviation signals driving further schema updates (Bein & Niv, 2023) or selectively activating other schemas through pattern separation (Figure 3C). This enables affective interaction states between different individuals to shift from negative (Figure 3B) to positive (Figure 3D) through changes in activated schemas.

In summary, the Affective Schema Model provides a comprehensive framework for understanding how biological agents utilize learned affective responses in combination with innate neural mechanisms. The model emphasizes the importance of associative learning, even in organisms with simple nervous systems, and the roles of conditioning and stimulus generalization in shaping affective responses. By forming affective schemas, organisms can efficiently navigate their environment, responding to novel stimuli based on past experiences and pattern recognition. This model not only highlights the complexity and adaptability of affective responses across species but also underscores the importance of cortical and subcortical pathways in processing affective stimuli. Consequently, the Affective Schema Model offers valuable insights into the complex mechanisms underlying affective interactions and their evolutionary significance, bridging the gap between simple neural responses and complex affective processing. This understanding paves the way for further exploration in fields like affective computing, where mimicking these biological processes may yield more sophisticated and responsive artificial systems.

[Figure 3: see original paper]

3.4 Comparison and Validation of Affective Interaction Models

In this section, we compare our proposed affective interaction models with existing affective decision-making and cognitive communication models, exploring their conceptual and functional similarities and differences (see Table 1).

Similar to the socially enactive cognitive system described by Kahl and Kopp (2023), our affective interaction models also contain two circulating systems of information flow: information flow within each agent and external information flow between two similar agents. Friston and Frith (2015) reframe theory of mind, emphasizing the inference of others' behavior through internal generative models. In affective interaction, this means effective communication requires both parties to possess sufficiently similar affective action models, enabling one to use its internal model to infer the other's affective intentions.

A fundamental prerequisite for cross-species affective interaction may be that different individuals possess similar internal affective action models, with some core affects serving as a "code library" for interaction. These core affects are fundamental and closely related to survival. Human affect research divides emotion into different components, including subjective self-reports and changes in behavioral, physiological, and neural activities. Applying these objective measurement methods to animal affect research not only avoids debates about animal

consciousness but also captures relationships between affect and bodily actions or changes. Therefore, affect can be operationally defined in behavioral terms as states triggered by rewards or punishments (Leknes & Tracey, 2008; Mendl & Paul, 2020). Affective valence and behavioral activation divide these core affects into dimensions like positive/negative and approach/avoidance. Mendl and Paul (2020) add arousal as a dimension to represent affect's impact on action vitality, defining four core affective states.

Loewenstein and Lerner's (2003) human affective decision-making model and Mendl and Paul's (2020) animal affective decision-making model both emphasize affect's influence on behavior from two perspectives: the impact of current affective states on decision-making, and the affect (i.e., reward or punishment) expected from future action outcomes. For the former, this influence primarily manifests in detecting new states and goals, subsequently affecting organisms' behavioral choices. Notably, the interaction between sensation and affect is bidirectional; sensation is not only influenced by affect but also serves as a tool for affect regulation (Rodriguez & Kross, 2023). All animals rapidly detect external environmental information through sensation, a prerequisite for adaptation and interaction. Environmental stimuli activate receptors throughout the body, converting physical energies like sound, light, electricity, and heat into neural impulses. These electrical signals generated by the body transmit through the central nervous system to the brain, where they are organized, processed, and interpreted. The objective external world connects with, is perceived by, and understood by biological entities through the sensory system, with the body as this process's starting point. Neuroscience research reveals sensory information's activating effect on affective responses, with neural signals rapidly entering sensory cortex and directly projecting to affect-generating brain regions like the amygdala, insula, and orbitofrontal cortex (Koelsch, 2018; Sullivan et al., 2015; Veldhuizen et al., 2020). Consequently, sensory stimuli are endowed with affective coloration (whether stimuli are pleasant or unpleasant), driving individuals to produce approach or avoidance behavioral responses. Unlike Mendl and Paul (2020), who treat internal states and core affective states as two separate factors influencing perception and memory, we believe sensory information provided by the body can integrate and reach certain thresholds to trigger higher-level affective responses or change current mood states. For example, in early mother-infant interactions, maternal touch can soothe infants' emotions, making them feel safe and loved (Su & Su, 2018); similar phenomena of touch-induced pleasure have been observed in non-human primates (Grandi & Gerbella, 2016).

To validate ATM across species, psychophysical experiments can be designed to capture and measure cross-species affective responses under controlled stimulus conditions. This validation plan includes adaptive subjective self-report methods for human participants and similar behavioral observations for non-human animals, such as rodents' ultrasonic vocalizations or specific postural changes in other species. Additionally, physiological sensors can objectively record responses, providing a dual-measurement approach to improve data reliability.

Each participating species can be exposed to a series of specially designed affective stimuli that trigger directly comparable responses across species. This approach not only tests the model's general applicability to affective threshold predictions but also explores nuanced differences in how different species perceive and respond to similar affective stimuli. By integrating behavioral and physiological data, we can verify the model's accuracy in predicting affective interactions across various biological agents, thereby ensuring the model reflects the true complexity of cross-species affective dynamics. This comprehensive approach aims to solidify the model's foundation in affective science and enhance its practical utility in real-world cross-species applications.

In the Dynamic Set-Point Model (Figure 2A), sensory signals from inside and outside the body transmit upward, comparing with dynamic internal homeostatic goals (a set of physical or biochemical states and psychological or affective states that biological entities need to maintain). Negative feedback regulation achieves organismic control. This process enables assessment and judgment of bodily states, with affect generated as a representational signal. These signals can undergo cognitive planning and management in pathways above threshold, or rapid activation responses in pathways below threshold. Since this model assumes adaptive cross-species interaction based on environmental feedback, we propose a comprehensive method including behavioral experiments, physiological monitoring, and computational modeling for validation. Control experiments can observe how different species adjust behavioral and physiological responses under different environmental conditions, using wearable biosensors to map neurophysiological activities related to affective processing and autonomic regulation. Meanwhile, agent-based computational simulations can model dynamic adjustments of homeostatic goals to predict interaction patterns, confirming the model's accuracy across various species and ecological niches by correlating experimental data with simulation results, thereby providing comprehensive validation.

During action organization, Mendl and Paul (2020) argue that animals compare current situations with similar remembered situations before making specific decisions. Retrieved memories contain previously taken actions in similar situations and their outcomes (punishment or reward), with animals ultimately choosing actions that generate the most positive affect. This bottom-up matching must undergo cognitive processes like perception, memory, and comparison to activate corresponding actions, which is not conducive to adapting to rapidly changing environments. Moreover, this model does not adequately explain the specific mechanism for matching situations to memories, thus failing to fully illustrate how animals can rapidly act based on past experiences when facing novel stimuli. Our proposed Affective Schema Model (Figure 3A) attempts to better explain such situations. We introduce the concept of schemas—schemas are the result of individuals' entire life experiences, including both universally present low-level stimulus generalization and socially transmitted cultural connotations. Different schemas have hierarchical relationships, with specific schema activation based on speculation and completion of entire events from contex-

tual cues. Even without truly recalling specific events, schemas formed from past experiences unconsciously influence our predictions about current situation development trends and guide animals to respond before changes occur. For example, animals may fear unfamiliar large objects even when not currently harmed. This anticipatory strategy is crucial for any animal's survival. Since this model assumes dynamic cross-species interaction based on learned affective responses, validating it requires a comprehensive approach combining conditioning behavior paradigms, neuroimaging, and computational simulation. Initial behavioral experiments across multiple species can track affective schema activation and adaptation under controlled stimuli, assessing changes from negative to positive behaviors as schemas evolve. Neuroimaging techniques like functional magnetic resonance imaging (fMRI) can visualize brain regions involved in schema processing, linking physiological data to observed behaviors. Additionally, computational models can simulate these interactions, predicting and further analyzing behaviors under different conditions to enhance model robustness. Combining these methods can rigorously test the model's ability to accurately describe affective learning and adaptation across different biological contexts.

4. Applying Affective Interaction Models to AI-Human Interaction Design

In developing AI capabilities, the challenge lies not merely in mimicking cognition but in encompassing the comprehensive human experience. Affective communication, based on shared evolutionary history and survival needs, offers a pathway to bridge this gap. Drawing inspiration from nature's rich affective exchanges, this paper aims to explore design principles for AI that can resonate affectively with humans and other forms of biological intelligence. Effective AI-human interaction should seamlessly combine cognitive understanding with affective resonance, ensuring interactions are natural, authentic, and impactful. Leveraging the potential of the proposed Large Affective Model, this paper aims to create interactive interfaces that can not only interpret user input but also simulate and convey affective states and goals, facilitating more comprehensive AI-human interactions.

4.1 The Necessity of Structured Models in AI-Human Affective Interaction

While evolutionary theory provides crucial evidence for understanding shared affective mechanisms, applying these to AI-human interaction requires structured frameworks. These models transform broad evolutionary concepts into concrete strategies for creating human-centered machine interfaces.

Pioneering research from the 1980s-90s used computational learning to model human-environment interactions (Jordan & Rumelhart, 1992; Wolpert et al., 1995). Early theoretical models employed stochastic generative models (Hinton

et al., 1995) and reinforcement learning (Barto et al., 1983), making important contributions to AI development by implicitly collecting world data. These models explored the physiological basis of internal models and proposed that internal representation creation could optimize AI-environment interactions (Doya, 1999; Imamizu et al., 2000).

A robust model should reduce algorithmic complexity according to the minimum description or message length principle (Wallace, 1999). This approach helps filter noise, improve prediction of real data, and enhance computational efficiency under resource constraints. For example, in neurobiology, model simplification often involves synaptic pruning or deleting redundant parameters (Tononi & Cirelli, 2006). From the basic M-P neuron model (McCulloch & Pitts, 1943) to advanced neural networks, the guiding principle is minimizing complexity. Artificial neural networks inspired by biological nervous system structures learn complex patterns through interconnected computational units, demonstrating the effectiveness of mimicking biological intelligence.

Darwin's evolutionary theory states that living systems thrive due to specific internal component arrangements and dynamic interactions (Cohen & Harel, 2006). Treating these arrangements as information (Cohen, 2006) allows us to view life's evolutionary trajectory as a balance between information increase and destruction. Entropy represents information, meaning complex arrangements are inherently unstable (Cohen, 2016). Therefore, biological evolution can be seen as an entropic selection process where arrangements that can resist entropic destructive effects are preserved.

The integration of biology and AI requires not merely replicating behavior and appearance through algorithms and data, but extracting universal internal models from genuine biological activities, grounded in phenomenological observation and integrated into learning algorithms. This ensures data is processed meaningfully, reflecting the nuanced complexity of biological intelligence.

4.2 Model Construction: Large Affective Model and Affective Interaction Models

Affective expression transcends language barriers, encompassing multimodal domains that integrate visual, auditory, olfactory, and tactile cues, along with more subtle signals like gait, movement trajectories, tear stains, eye contact, skin color, and body tremors (Ezzameli & Mahersia, 2023; Poria et al., 2017). Even in the visual domain alone, interpreting these diverse affective feature data presents enormous algorithmic challenges (Papadimitriou, 2020). Currently, affective computing research increasingly emphasizes using multimodal data to improve affective recognition capability, optimizing algorithms for richer multimodal interactions and reducing inherent dataset biases (Cortiñas-Lorenzo & Lacey, 2023; Ma & Yarosh, 2023; Yu et al., 2021). However, these efforts still lack deep evolutionary understanding of affective processes.

As previously discussed, different species demonstrate affective communication

abilities crucial for survival across various environmental interactions. This highlights the importance of developing a framework that can both analyze and simulate these complex affective dynamics. Schiller et al.'s (2024) "Human Affectome" provides a structured framework for dissecting and reorganizing complex affective phenomena, facilitating deeper understanding of affective dynamics. We believe this framework applies not only to analyzing human affective phenomena but can be extended to broader affective interactions in nature, as affect is equally significant for all organisms from a teleological perspective. Although affective phenomena are often intertwined with other dynamic processes in neurobiological systems (such as perception, attention, memory, and motor functions), these processes can be distilled into algorithms reflecting survival-driven behavior. Schiller et al. (2024) divide these processes into two groups: affective concern processes, which process physical or psychological objects—including things, people, situations, and representations of the past and predictions of the future—with the purpose of ensuring survival capacity, performing operations, establishing relevance, and abstraction (this function is mainly limited to human-like intelligence); and affective characteristic processes, which manage organisms' own experiences through adaptation, characterized by valence and arousal. Current affective computing research has significant data dimension limitations: while multimodal physiological characteristic data dominates mainstream research, indicators of affective concern reflecting organisms' behavioral tendencies are universally missing from data collection and interpretation systems. These indicators essentially reflect organisms' proactive behavioral orientations based on valence evaluation. According to the immediacy of action impact, relevance hierarchies can be organized from near to far, or global assessments can be made of the adaptability of different affective concerns over time (Schiller et al., 2024). We position our three proposed affective interaction models at different levels of affective concern to help establish complete affective data and algorithm structures (Figure 4 [Figure 4: see original paper]).

[Figure 4: see original paper]

Concerns about organisms' physiological states represent the most direct and fundamental level of affective concern, typically requiring direct and concrete actions to address. Environmental stimuli trigger a series of interoceptive sensations, and the integration of these internal bodily signals constitutes current affective experience. This concern group is closely related to survival goals; therefore, a critical physiological affective experience is the sense of safety. We link the Affective Threshold Model to these physiological-level, immediate affective concerns, suggesting that algorithms addressing this issue can draw inspiration from functional relationships between physical quantities and psychophysical quantities in sensation research. Mathematical models established between external stimuli and resulting internal sensations can effectively predict action tendencies toward objects. Physiological affective concerns can be quantified as a set of values that make organisms feel safe and comfortable, including temperature, energy levels, nutritional status, etc. Therefore, affective data needed to enhance AI-human interaction should include both the physical quantity of

stimuli and the intensity of resulting sensations.

Beyond immediate and direct physiological concerns, more advanced organisms typically require interaction with the environment over longer timeframes through a series of more complex actions to ensure vitality in future environments, which Schiller et al. (2024) term "operational concerns." Notably, to comprehensively address multidimensional problems facing these organisms as completely as possible, these operational concerns are both diverse and complex. Although they can be categorized into cooperation, morality, and aesthetic concerns, a major challenge in affective modeling is simulating the weight changes of these affective concerns across different moments. To address this, we explain this process through dynamic adjustment of organisms' internal homeostatic goals. We argue that biological organisms' internal homeostatic goals are not merely constant physiological state values but are sets of data structures that adjust with environmental changes, encompassing certain more abstract affective needs that organisms must maintain, such as cooperation needs, aesthetic appreciation, and moral values (similar to "operational concerns"). A feasible method for achieving weight changes is conducting Bayesian analysis of the probability of generating specific homeostatic goals when objects appear in the environment, using probability to represent changes in relevance gradients. This algorithm is similar to performing "spell check," identifying the most likely correct spelling from several known error options. When applied to homeostatic goal computation, this means affective AI systems need to perceive all objects in the environment potentially relevant to users and calculate the most likely goals in that context.

At the global level, affective concerns are no longer driven by specific objects but integrate concerns at all levels into an overall state (LeDoux, 2012; Schiller et al., 2024). The Affective Schema Model reflects this abstraction and generalization of affective phenomena over time. Schema formation originates from individuals' entire experiential history with environmental interaction. During information "compression," specificity related to particular contexts disappears, aggregating into schemas characterized by two affective valence features: positive affective schemas and negative affective schemas. Stimuli activating relevant schemas are no longer concrete, explicit objects but partial cues such as features, outlines, smells, and sounds. These affective schemas enable agents to rapidly adapt to environments lacking specific information.

Therefore, affective AI systems must not only identify users' current affective states but also "understand" the associations between these external affective expressions and their deep psychological models—truly "understanding" users' affective concerns and their expression in contexts displayed as affective characteristics. In this context, we anticipate that in affective computing, a specialized "Large Affective Model" (LAM) will be essential for addressing highly contextualized alignment problems in multimodal data—challenges that traditional large language models may struggle to handle. This model's uniqueness lies in its ability to deeply understand, interpret, and generate multimodal informa-

tion related to various dimensions of affect in specific contexts. It leverages extensive sensory modalities such as gaze, touch, tone of voice, taste, smell, balance, and body movement, integrating these cues within highly contextualized frameworks. This approach treats affect as an effective response to stimuli characterized by arousal and valence, based on consistency in cross-species affective mechanisms, while considering differences in affective expression and interactions between primary affective networks and higher cognitive functions in the brain (Panksepp, 2012).

The Large Affective Model (LAM) works synergistically with our three affective interaction models derived from empirical observations of interspecies affective communication. These models are not merely mechanical replications of behavior but are analyzed from genuine interaction characteristics of biological entities. They integrate the essence of affect into computational processes, enriching the depth and relevance of AI-human interactions.

- **Affective Threshold Model (ATM):** This model investigates individuals' sensitivity and response thresholds to environmental stimuli, elucidating the interaction between affective states and stimuli. Mathematical models developed from this research mapping external stimuli to affect can address physiological-level affective concerns.
- **Dynamic Set-Point Model (DSPM):** This model focuses on monitoring and dynamically adjusting physiological and affective states based on environmental feedback, striving to maintain internal and external perceptions within optimal ranges. Changes in homeostatic goal weights across different scenarios reflect the most important operational-level affective concerns at present.
- **Affective Schema Model (ASM):** This model involves activating relevant affective schemas based on stimulus evaluation, thereby facilitating rapid behavioral responses. This reflects how affective concerns shift from reactions to specific objects to global states integrating different levels of concern.

In summary, we argue that affective data required for constructing Large Affective Models should include objects that trigger specific affective experiences. Furthermore, because affective concerns have hierarchical and global characteristics, Large Affective Models must be able to flexibly adjust affective algorithms according to contextual changes to more accurately predict and guide agents' behavior. It is important to note that our goal is not to design a super-anthropomorphic "competitor" that surpasses human performance in multitasking, but to create a "new species" capable of affective interaction with humans and other forms of biological intelligence.

[Figure 5: see original paper]

5. Advancing AI Alignment Through Affective Understanding

The rapid development of AI technology brings benefits alongside potential risks. Recent research reveals concerning trends in large language models (LLMs),

such as inherent biases and inaccuracies (Bang et al., 2023; Perez et al., 2022), and power-seeking behavioral tendencies (Si et al., 2022, pp. 2659-2673). Additionally, AI systems may be used for malicious purposes (Bubeck et al., 2023). Like the transformative impacts of industrial and digital revolutions, AI heralds a powerful new technological era that may significantly alter social dynamics. Guiding AI's integration into social structures in an ethical manner is crucial to ensure these changes promote equity and remain controllable across all social strata (Peeters et al., 2021; Walther, 2021). In this context, researchers including Ji et al. (2023) call for AI alignment research, emphasizing that AI system design must prioritize empathy for human emotional and psychological needs, enhance human welfare, respect privacy, and contribute positively to social progress. We believe that research and development of affective AI will promote AI alignment, equipping AI systems with advanced qualities as collaborative partners in human-AI coexistence ecosystems.

The foundational stage of AI alignment research aims to synchronize AI actions with user intentions. This includes accurately interpreting and executing direct instructions, intuitively anticipating users' unexpressed desires, and adaptively responding to users' behavioral cues (Gabriel, 2020; Ji et al., 2023). Currently, large language models excel at following explicit instructions—often demonstrated in dialogue-based interactions—but frequently fail to capture deeper human insights. This deficiency partly stems from inherent limitations of language-mediated interaction: deep-seated intentions are often closely linked to affect, representing individuals' preferences, aversions, and orientations that are not necessarily expressed through language. While language can indeed carry affect, and some scholars attempt to integrate multimodal information through highly expressive language to achieve semantic alignment across modalities (Zhu et al., 2023), this approach may distort the path to understanding affect. Affect is fundamentally an important indicator of organisms' interactions with their environment, directly influencing decision-making and action aimed at ensuring survival and improving adaptability. Recognizing this is crucial for AI alignment.

Therefore, realigning AI systems to understand and integrate richly contextualized human affective concerns, rather than merely responding to explicit ethical or value statements, may yield more genuinely beneficial AI outcomes for individuals and society, whether biological or artificial.

To advance this form of AI alignment, we advocate for developing and refining Large Affective Models (LAMs) tailored to individuals' and society's unique affective landscapes. By using diverse models to depict various dimensions and levels of affective concerns, we can facilitate truly context-aware interactions. This strategy not only improves the relevance and applicability of AI-human interactions but also enhances the transparency and interpretability of AI predictions, providing a powerful complement to traditional data-driven machine learning methods.

6. Conclusion: Connecting Insights from Biological Evolution to AI's Future

Since its birth at the 1956 Dartmouth Conference, AI has experienced four important developmental stages, each marked by unique advances and directions. The first three stages focused on symbolic rule-based reasoning, machine learning techniques, and probabilistic reasoning, laying the groundwork for exploring AI's foundational theories and technologies. The current fourth stage, characterized by deep learning and big data, signifies not only technological breakthroughs but also a shift toward exploring AI application scenarios and advancing toward Artificial General Intelligence (AGI).

On the path to AGI, the choice of direction is crucial. Cognitive science, long intertwined with computer science, has historically drawn analogies between human thinking and computers. However, focusing solely on computational strategies to replicate human cognition overlooks algorithms' inherent limitations in handling unstructured real-life problems. These limitations manifest in two key aspects: first, the difficulty of conveying true needs and desires to robots/agents through logical instructions; second, people's "algorithm aversion" to purely computationally derived results in situations requiring empathy (Dietvorst et al., 2015; Karataş & Cutright, 2023).

Transforming AGI from pure cognitive systems into affective entities through increased computational power and model scaling alone is insufficient. While large language models like GPT-4 excel at generating content in symbolic spaces, they cannot guarantee this content's authenticity. This raises concerns about their potential use in creating attack vectors targeting individuals (Bubeck et al., 2023). However, biological organisms like humans process information and interact with the world beyond symbolic spaces. The brain's neural network integrates various non-symbolic representations, achieving the fusion of symbolic computation and affective judgment. While affect can be symbolized, key information is lost in this process. Today's multimodal large models, primarily based on language models, still differ from biological thought logic. Endowing AI with affective capabilities enables it to generate meaning and develop more complex moral systems. This allows AI to transcend mere obedience to input commands, making judgments before execution and potentially refusing harmful or misleading information. This transformation helps address risks arising from AI's pure instrumentality.

Our three proposed affective interaction models represent analyses and summaries of affective activity characteristics across different intelligent entities, representing a realistically feasible affect-behavior logic. These models prioritize rapid activation through sub-threshold pathways and body-centered responses, avoiding the complexity of higher cognitive processes. From the perspective of biological evolution, they provide a universal viewpoint emphasizing that symbolic representation emerges later than high-level conceptual knowledge. This view aligns with recent shifts in cognitive science that advocate eliminating men-

tal representations from cognitive explanations (Chemero, 2013) and emphasize multidimensional analysis integrating body, brain, and environment (Menary, 2010).

In summary, in AI's developmental trajectory, connecting insights from biological evolution to AI's future requires balanced integration of cognitive capabilities and affective intelligence. This comprehensive approach promises a more nuanced, ethical, and human-centered path for AI development, fostering a future where technology not only complements but also enriches human experience.

Author Contributions
Chongyi Liu: Writing—original draft, Visualization, Investigation, Formal analysis, Conceptualization.
Bin Yin: Writing—review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Conflict of Interest Statement
The authors declare they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement
The research described in this article did not generate any new data.

Funding
This research was supported by the Joint Fund for Cross-Strait Scientific and Technological Cooperation (Grant No. U1805263) and the Humanities and Social Sciences Fund of the Ministry of Education (Grant No. 23YJAZH183). The funders were not involved in any aspect of the research preparation.

References

Abdollahi, H., Mahoor, M. H., Zandie, R., Siewierski, J., & Qualls, S. H. (2023). Artificial Emotional Intelligence in Socially Assistive Robots for Older Adults: A Pilot Study. *IEEE Transactions on Affective Computing*, *14*(3), 1050–1061. https://doi.org/10.1109/TAFFC.2022.3143803

Adolph, K. E. (2005). Learning to learn in the development of action. In *Action as an Organizer of Learning and Development* (pp. 91–122). Psychology Press.

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *International Conference on Machine Learning*.

AI4Science, M. R., & Quantum, M. A. (2023). The Impact of Large Language Models on Scientific Discovery: A Preliminary Study using GPT-4. *arXiv preprint arXiv:2311.07361*.

Akre, K. L., & Johnsen, S. (2014). Psychophysics and the evolution of behavior. *Trends in Ecology & Evolution*, *29*(5), 291–300.

https://doi.org/10.1016/j.tree.2014.03.007

Aminah, S., Hidayah, N., & Ramli, M. (2023). Considering ChatGPT to be the first aid for young adults on mental health issues. *Journal of Public Health*, *45*(3), e615–e616. https://doi.org/10.1093/pubmed/fdad065

Arbib, M. A., & Fellous, J.-M. (2004). Emotions: From brain to robot. *Trends in Cognitive Sciences*, *8*(12), 554–561. https://doi.org/10.1016/j.tics.2004.10.004

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064. https://doi.org/10.1038/s41562-017-0064

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349. https://doi.org/10.1016/j.cognition.2009.07.005

Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *arXiv preprint arXiv:2302.04023*.

Barker, S. B., Knisely, J. S., Schubert, C. M., Green, J. D., & Ameringer, S. (2015). The Effect of an Animal-Assisted Intervention on Anxiety and Pain in Hospitalized Children. *Anthrozoös*, *28*(1), 101–112. https://doi.org/10.2752/089279315X14129350722091

Bartal, I. B.-A., Decety, J., & Mason, P. (2011). Empathy and pro-social behavior in rats. *Science*, *334*(6061), 1427–1430. https://doi.org/10.1126/science.1210789

Bartal, I. B.-A., Shan, H., Molasky, N. M., Murray, T. M., Williams, J. Z., Decety, J., & Mason, P. (2016). Anxiolytic treatment impairs helping behavior in rats. *Frontiers in Psychology*, *7*, 850. https://doi.org/10.3389/fpsyg.2016.00850

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-13*(5), 834–846. https://doi.org/10.1109/TSMC.1983.6313077

Bein, O., & Niv, Y. (2023). Schemas, reinforcement learning, and the medial prefrontal cortex. *PsyArXiv*. https://doi.org/10.31234/osf.io/spxq9

Berkovich, R., & Meiran, N. (2023). Pleasant emotional feelings follow one of the most basic psychophysical laws (Weber's law) as most sensations do. *Emotion*, *23*(5), 1213–1223. https://doi.org/10.1037/emo0001161

Bielecki, J., Dam Nielsen, S. K., Nachman, G., & Garm, A. (2023). Associative learning in the box jellyfish *Tripedalia cystophora*. *Current Biology*. https://doi.org/10.1016/j.cub.2023.08.056

Binz, M., Dasgupta, I., Jagadish, A. K., Botvinick, M., Wang, J. X., & Schulz, E. (2023). Meta-learned models of cognition. *Behavioral and Brain Sciences*, *46*, e1. https://doi.org/10.1017/S0140525X23003266

Broadbent, E., Kumar, V., Li, X., Sollers, J., 3rd, Stafford, R. Q., MacDonald, B. A., & Wegner, D. M. (2013). Robots with Display Screens: A Robot with a More Humanlike Face Display Is Perceived To Have More Mind and a Better Personality. *PLOS ONE*, *8*(8), e72589. https://doi.org/10.1371/journal.pone.0072589

Broekens, J., Hilpert, B., Verberne, S., Baraka, K., Gebhard, P., & Plaat, A. (2023, September). Fine-grained Affective Processing Capabilities Emerging from Large Language Models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 1-8). IEEE. https://doi.org/10.1109/ACII57435.2023.10361618

Brosschot, J. F., Verkuil, B., & Thayer, J. F. (2016). The default response to uncertainty and the importance of perceived safety in anxiety and stress: An evolution-theoretical perspective. *Journal of Anxiety Disorders*, *41*, 22–34. https://doi.org/10.1016/j.janxdis.2016.04.012

Brown, S. P. (2006). Cooperation: Integrating evolutionary and ecological perspectives. *Current Biology*, *16*(22), R960–R961. https://doi.org/10.1016/j.cub.2006.10.019

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Craig, A. D. (B). (2003). Interoception: The sense of the physiological condition of the body. *Current Opinion in Neurobiology*, *13*(4), 500–505. https://doi.org/10.1016/S0959-4388(03)00090-4

Carleton, R. N. (2012). The intolerance of uncertainty construct in the context of anxiety disorders: Theoretical and practical perspectives. *Expert Review of Neurotherapeutics*, *12*(8), 937–947. https://doi.org/10.1586/ern.12.82

Chang, L. J., Gianaros, P. J., Manuck, S. B., Krishnan, A., & Wager, T. D. (2015). A Sensitive and Specific Neural Signature for Picture-Induced Negative Affect. *PLOS Biology*, *13*(6), e1002180. https://doi.org/10.1371/journal.pbio.1002180

Chemero, A. (2013). *Radical Embodied Cognitive Science*. Review of General Psychology, *17*(2), 145–150. https://doi.org/10.1037/a0032923

Clark, H. H., & Fischer, K. (2023). Social robots as depictions of social agents. *Behavioral and Brain Sciences*, *46*, e21. https://doi.org/10.1017/S0140525X22000668

Cohen, I. R. (2006). Informational Landscapes in Art, Science, and Evolution. *Bulletin of Mathematical Biology*, *68*(5), 1213–1229. https://doi.org/10.1007/s11538-006-9118-4

Cohen, I. R. (2016). Updating Darwin: Information and entropy drive the evolution of life. *F1000Research*, *5*, 2808. https://doi.org/10.12688/f1000research.10289.1

Cohen, I. R., & Harel, D. (2006). Explaining a complex living system: Dynamics, multi-scaling and emergence. *Journal of the Royal Society Interface*, *3*(13),

963–970. https://doi.org/10.1098/rsif.2006.0173

Cortiñas-Lorenzo, K., & Lacey, G. (2023). Toward Explainable Affective Computing: A Review. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(9), 5432–5445. https://doi.org/10.1109/TNNLS.2023.3270027

Critchley, H. D., & Garfinkel, S. N. (2017). Interoception and emotion. *Current Opinion in Psychology*, *17*, 7–14. https://doi.org/10.1016/j.copsyc.2017.04.020

Cromwell, H. C., & Panksepp, J. (2011). Rethinking the cognitive revolution from a neural perspective: How overuse/misuse of the term 'cognition' and the neglect of affective controls in behavioral neuroscience could be delaying progress in understanding the BrainMind. *Neuroscience & Biobehavioral Reviews*, *35*(9), 2021–2035. https://doi.org/10.1016/j.neubiorev.2011.02.008

Damasio, A., & Carvalho, G. B. (2013). The nature of feelings: Evolutionary and neurobiological origins. *Nature Reviews Neuroscience*, *14*(2), 143–152. https://doi.org/10.1038/nrn3403

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*, *12*(7), 961–974. https://doi.org/10.1016/S0893-6080(99)00046-5

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, *55*(9). https://doi.org/10.1145/3561048

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When We Need A Human: Motivational determinants of anthropomorphism. *Social Cognition*, *26*(2), 143–155. https://doi.org/10.1521/soco.2008.26.2.143

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Ezzameli, K., & Mahersia, H. (2023). Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion*, *99*, 101847. https://doi.org/10.1016/j.inffus.2023.101847

Friston, K., & Frith, C. (2015). A Duet for one. *Consciousness and Cognition*, *36*, 390–405. https://doi.org/10.1016/j.concog.2014.12.003

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Godinho, F., Magnin, M., Frot, M., Perchet, C., & Garcia-Larrea, L. (2006). Emotional Modulation of Pain: Is It the Sensation or

What We Recall? *The Journal of Neuroscience*, *26*(44), 11454–11461. https://doi.org/10.1523/JNEUROSCI.2260-06.2006

Grandi, L. C., & Gerbella, M. (2016). Single Neurons in the Insular Cortex of a Macaque Monkey Respond to Skin Brushing: Preliminary Data of the Possible Representation of Pleasant Touch. *Frontiers in Behavioral Neuroscience*, *10*. https://doi.org/10.3389/fnbeh.2016.00090

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, *56*(1), 51–65. https://doi.org/10.1037/h0062474

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

Hauser, M., McAuliffe, K., & Blake, P. R. (2009). Evolving the ingredients for reciprocity and spite. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1533), 3255–3266. https://doi.org/10.1098/rstb.2009.0116

Heller, A. S. (2020). From Conditioning to Emotion: Translating Animal Models of Learning to Human Psychopathology. *The Neuroscientist*, *26*(1), 107–116. https://doi.org/10.1177/1073858419866820

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995). The "Wake-Sleep" Algorithm for Unsupervised Neural Networks. *Science*, *268*(5214), 1158–1161. https://doi.org/10.1126/science.7761831

Houlihan, S. D., Kleiman-Weiner, M., Hewitt, L. B., Tenenbaum, J. B., & Saxe, R. (2023). Emotion prediction as computation over a generative theory of mind. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *381*(2251), 20220047. https://doi.org/10.1098/rsta.2022.0047

Huzard, D., Martin, M., Maingret, F., Chemin, J., Jeanneteau, F., Mery, P.-F., Fossat, P., Bourinet, E., & François, A. (2022). The impact of C-tactile low-threshold mechanoreceptors on affective touch and social interactions in mice. *Science Advances*, *8*(26), eabo7566. https://doi.org/10.1126/sciadv.abo7566

Hyett, M., Parker, G., & Breakspear, M. (2014). Bias and discriminability during emotional signal detection in melancholic depression. *BMC Psychiatry*, *14*(1), 122. https://doi.org/10.1186/1471-244X-14-122

Imamizu, H., Miyauchi, S., Tamada, T., Sasaki, Y., Takino, R., Pütz, B., Yoshioka, T., & Kawato, M. (2000). Human cerebellar activity reflecting an acquired internal model of a new tool. *Nature*, *403*(6766), 192–195. https://doi.org/10.1038/35003194

Izard, C. E. (2009). Emotion Theory and Research: Highlights, Unanswered Questions, and Emerging Issues. *Annual Review of Psychology*, *60*(1), 1–25. https://doi.org/10.1146/annurev.psych.60.110707.163539

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K. Y., Dai, J., Pan, X., O'Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., … Gao, W. (2023). AI Alignment: A Comprehensive Survey. *arXiv preprint arXiv:2310.19852*.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, *16*(3), 307–354. https://doi.org/10.1016/0364-0213(92)90036-T

Kahl, S., & Kopp, S. (2023). Intertwining the social and the cognitive loops: Socially enactive cognition for human-compatible interactive systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1875), 20210474. https://doi.org/10.1098/rstb.2021.0474

Karataş, M., & Cutright, K. M. (2023). Thinking about God increases acceptance of artificial intelligence in decision-making. *Proceedings of the National Academy of Sciences*, *120*(33), e2218961120. https://doi.org/10.1073/pnas.2218961120

Karmon-Presser, A., Sheppes, G., & Meiran, N. (2018). How does it "feel"? A signal detection approach to feeling generation. *Emotion*, *18*(1), 94–115. https://doi.org/10.1037/emo0000298

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, *34*(7), 1185–1243. https://doi.org/10.1111/j.1551-6709.2010.01128.x

Keysers, C., Knapska, E., Moita, M. A., & Gazzola, V. (2022). Emotional contagion and prosocial behavior in rodents. *Trends in Cognitive Sciences*, *26*(8), 727–742. https://doi.org/10.1016/j.tics.2022.05.005

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

Koelsch, S. (2018). Investigating the Neural Encoding of Emotion with Music. *Neuron*, *98*(6), 1075–1079. https://doi.org/10.1016/j.neuron.2018.04.029

Krusemark, E. A., & Li, W. (2011). Do All Threats Work the Same Way? Divergent Effects of Fear and Disgust on Sensory Perception and Attention. *The Journal of Neuroscience*, *31*(9), 3429–3434. https://doi.org/10.1523/JNEUROSCI.4394-10.2011

Lazarus, R. S. (1984). On the primacy of cognition. *American Psychologist*, *39*(2), 124–129. https://doi.org/10.1037/0003-066X.39.2.124

Leavens, D. A., Russell, J. L., & Hopkins, W. D. (2005). Intentionality as Measured in the Persistence and Elaboration of Communication by Chimpanzees (*Pan troglodytes*). *Child Development*, *76*(1), 291–306. https://doi.org/10.1111/j.1467-8624.2005.00845.x

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeDoux, J. E. (2012). Rethinking the Emotional Brain. *Neuron*, *73*(4), 653–676. https://doi.org/10.1016/j.neuron.2012.02.004

Leknes, S., & Tracey, I. (2008). A common neurobiology for pain and pleasure. *Nature Reviews Neuroscience*, *9*(4), 314–320. https://doi.org/10.1038/nrn2333

Li, X., Li, Y., Joty, S., Liu, L., Huang, F., Qiu, L., & Bing, L. (2023). Does GPT-3 Demonstrate Psychopathy? Evaluating Large Language Models from a Psychological Perspective. *arXiv preprint arXiv:2303.12566*.

Li, X., & Sung, Y. (2021). Anthropomorphism brings us closer: The mediating role of psychological distance in User–AI assistant interactions. *Computers in Human Behavior*, *118*, 106680. https://doi.org/10.1016/j.chb.2021.106680

Loewenstein, G., & Lerner, J. S. (2003). The role of affect in decision making. In *Handbook of Affective Science* (Vol. 619, p. 3). Oxford University Press.

Łopuch, S., & Popik, P. (2011). Cooperative behavior of laboratory rats (*Rattus norvegicus*) in an instrumental task. *Journal of Comparative Psychology*, *125*(2), 250–253. https://doi.org/10.1037/a0022389

Ma, H., & Yarosh, S. (2023). A Review of Affective Computing Research Based on Function-Component-Representation Framework. *IEEE Transactions on Affective Computing*, *14*(2), 1655–1674. https://doi.org/10.1109/TAFFC.2021.3104512

Mamassian, P. (2016). Visual Confidence. *Annual Review of Vision Science*, *2*(1), 459–481. https://doi.org/10.1146/annurev-vision-111815-114630

Marr, D., & Poggio, T. (1976). From understanding computation to understanding neural circuitry. *MIT AI Memo*, *357*.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

McShea, D. W. (2017). Logic, passion and the problem of convergence. *Interface Focus*, *7*(3), 20160122. https://doi.org/10.1098/rsfs.2016.0122

McShea, D. W. (2023). Evolutionary trends and goal directedness. *Synthese*, *201*(5), 178. https://doi.org/10.1007/s11229-023-04164-9

Menary, R. (2010). Introduction to the special issue on 4E cognition. *Phenomenology and the Cognitive Sciences*, *9*(4), 459–463. https://doi.org/10.1007/s11097-010-9179-9

Mendl, M., & Paul, E. S. (2020). Animal affect and decision-making. *Neuroscience & Biobehavioral Reviews*, *112*, 144–163. https://doi.org/10.1016/j.neubiorev.2020.01.025

Merkies, K., Crouchman, E., & Belliveau, H. (2022). Human Ability to Determine Affective States in Domestic Horse Whinnies. *Anthrozoös*, *35*(3), 345–358.

https://doi.org/10.1080/08927936.2021.1999605

Miller, E. K., & Cohen, J. D. (2001). An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, *24*(1), 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Minta, S. C., Minta, K. A., & Lott, D. F. (1992). Hunting Associations between Badgers (*Taxidea taxus*) and Coyotes (*Canis latrans*). *Journal of Mammalogy*, *73*(4), 814–820. https://doi.org/10.2307/1382201

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent Models of Visual Attention. *Advances in Neural Information Processing Systems*, 27. https://proceedings.neurips.cc/paper/2014/hash/09c6c3783b4a70054da74f2538ed47c6-Abstract.html

Moore, R. (2016). Meaning and ostension in great ape gestural communication. *Animal Cognition*, *19*(1), 223–231. https://doi.org/10.1007/s10071-015-0905-x

Morris, M. R., Sohl-dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., & Legg, S. (2023). Levels of AGI: Operationalizing Progress on the Path to AGI. *arXiv preprint arXiv:2311.02462*.

Morrissey, M. D., Insel, N., & Takehara-Nishiuchi, K. (2017). Generalizable knowledge outweighs incidental details in prefrontal ensemble code over time. *eLife*, *6*, e22177. https://doi.org/10.7554/eLife.22177

Mossio, M., & Bich, L. (2017). What makes biological organisation teleological? *Synthese*, *194*(4), 1089–1114. https://doi.org/10.1007/s11229-014-0594-z

Ngo, C. T., Michelmann, S., Olson, I. R., & Newcombe, N. S. (2021). Pattern separation and pattern completion: Behaviorally separable processes? *Memory & Cognition*, *49*(1), 193–205. https://doi.org/10.3758/s13421-020-01072-y

Nielsen, L., & Kaszniak, A. W. (2006). Awareness of subtle emotional feelings: A comparison of long-term meditators and nonmeditators. *Emotion*, *6*(3), 392–405. https://doi.org/10.1037/1528-3542.6.3.392

Nowak, M. A. (2012). Evolving cooperation. *Journal of Theoretical Biology*, *299*, 1–8. https://doi.org/10.1016/j.jtbi.2012.01.014

O'Madagain, C., & Tomasello, M. (2022). Shared intentionality, reason-giving and the evolution of human culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1843), 20200320. https://doi.org/10.1098/rstb.2020.0320

Ong, D. C., Soh, H., Zaki, J., & Goodman, N. D. (2021). Applying Probabilistic Programming to Affective Computing. *IEEE Transactions on Affective Computing*, *12*(2), 306–317. https://doi.org/10.1109/TAFFC.2019.2905211

Ong, D. C., Zaki, J., & Goodman, N. D. (2015). Affective cognition: Exploring lay theories of emotion. *Cognition*, *143*, 141–162. https://doi.org/10.1016/j.cognition.2015.06.010

Ong, D. C., Zaki, J., & Goodman, N. D. (2016). Emotions in lay explanations of behavior. *Proceedings of the Thirty-Eighth Annual Conference of the Cognitive Science Society*.

Ong, D. C., Zaki, J., & Goodman, N. D. (2019). Computational Models of Emotion Inference in Theory of Mind: A Review and Roadmap. *Topics in Cognitive Science*, *11*(2), 338–357. https://doi.org/10.1111/tops.12371

Panksepp, J. (2012). What is an emotional feeling? Lessons about affective origins from cross-species neuroscience. *Motivation and Emotion*, *36*(1), 4–15. https://doi.org/10.1007/s11031-011-9232-y

Papadimitriou, F. (2020). *Spatial Complexity: Theory, Mathematical Methods and Applications*. Springer International Publishing. https://doi.org/10.1007/978-3-030-59671-2

Peeters, M. M. M., van Diggelen, J., van den Bosch, K., Bronkhorst, A., Neerincx, M. A., Schraagen, J. M., & Raaijmakers, S. (2021). Hybrid collective intelligence in a human–AI society. *AI & SOCIETY*, *36*(1), 217–238. https://doi.org/10.1007/s00146-020-01005-y

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., … Kaplan, J. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. *arXiv preprint arXiv:2212.09251*.

Picard, R. W. (2000). *Affective Computing*. MIT Press.

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003

Rao, H., Leung, C., & Miao, C. (2023). Can ChatGPT Assess Human Personalities? A General Evaluation Framework. *arXiv preprint arXiv:2303.12666*.

Rodriguez, M., & Kross, E. (2023). Sensory emotion regulation. *Trends in Cognitive Sciences*, *27*(4), 379–390. https://doi.org/10.1016/j.tics.2023.01.008

Roli, A., Jaeger, J., & Kauffman, S. A. (2022). How organisms come to know the world: Fundamental limits on artificial general intelligence. *Frontiers in Ecology and Evolution*, *9*, 787726. https://doi.org/10.3389/fevo.2021.787726

Ruijten, P. A., Haans, A., Ham, J., & Midden, C. J. (2019). Perceived human-likeness of social robots: Testing the Rasch model as a method for measuring anthropomorphism. *International Journal of Social Robotics*, *11*, 477–494. https://doi.org/10.1007/s12369-019-00518-8

Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. *AJOB Neuroscience*, *11*(2), 88–95. https://doi.org/10.1080/21507740.2020.1740350

Saxe, R., & Houlihan, S. D. (2017). Formalizing emotion concepts within a Bayesian model of theory of mind. *Emotion, Stress, and Health*, *36*, 15–21. https://doi.org/10.1016/j.concog.2017.04.019

Schiller, D., Yu, A. N. C., Alia-Klein, N., Becker, S., Cromwell, H. C., Dolcos, F., Eslinger, P. J., Frewen, P., Kemp, A. H., Pace-Schott, E. F., Raber, J., Silton, R. L., Stefanova, E., Williams, J. H. G., Abe, N., Aghajani, M., Albrecht, F., Alexander, R., Anders, S., … Lowe, L. (2024). The Human Affectome. *Neuroscience & Biobehavioral Reviews*, *105450*. https://doi.org/10.1016/j.neubiorev.2023.105450

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schueller, S. M., & Morris, R. R. (2023). Clinical science and practice in the age of large language models and generative artificial intelligence. *Journal of Consulting and Clinical Psychology*, *91*(10), 559–561. https://doi.org/10.1037/ccp0000848

Schulkin, J., & Sterling, P. (2019). Allostasis: A Brain-Centered, Predictive Mode of Physiological Regulation. *Trends in Neurosciences*, *42*(10), 740–752. https://doi.org/10.1016/j.tins.2019.07.010

Scott-Phillips, T. C. (2015). Nonhuman Primate Communication, Pragmatics, and the Origins of Language. *Current Anthropology*, *56*(1), 56–80. https://doi.org/10.1086/679674

Si, W. M., Backes, M., Blackburn, J., De Cristofaro, E., Stringhini, G., Zannettou, S., & Zhang, Y. (2022). Why So Toxic? Measuring and Triggering Toxic Behavior in Open-Domain Chatbots. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2659–2673. https://doi.org/10.1145/3548606.3560599

Simen, P., Vlasov, K., & Papadakis, S. (2016). Scale (in)variance in a unified diffusion model of decision making. *Psychological Review*, *123*(2), 151–181. https://doi.org/10.1037/rev0000014

Slaby, J., & Stephan, A. (2008). Affective intentionality and self-consciousness. *Consciousness and Cognition*, *17*(2), 506–513. https://doi.org/10.1016/j.concog.2008.03.007

Sokolov, E. N., & Boucsein, W. (2000). A psychophysiological model of emotion space. *Integrative Physiological and Behavioral Science*, *35*(2), 81–119. https://doi.org/10.1007/BF02688770

Su, J., & Su, Y. (2018). A touch-scaffolded model of human prosociality. *Neuroscience & Biobehavioral Reviews*, *92*, 453–463. https://doi.org/10.1016/j.neubiorev.2018.07.008

Sullivan, R. M., Wilson, D. A., Ravel, N., & Mouly, A.-M. (2015). Olfactory memory networks: From emotional learning to social behaviors. *Frontiers in Behavioral Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00036

Tononi, G., & Cirelli, C. (2006). Sleep function and synaptic homeostasis. *Sleep Medicine Reviews*, *10*(1), 49–62. https://doi.org/10.1016/j.smrv.2005.05.002

Townsend, S. W., Koski, S. E., Byrne, R. W., Slocombe, K. E., Bickel, B., Boeckle, M., Braga Goncalves, I., Burkart, J. M., Flower, T., Gaunet, F., Glock, H. J., Gruber, T., Jansen, D. A. W. A. M., Liebal, K., Linke, A., Miklósi, Á., Moore, R., van Schaik, C. P., Stoll, S., … Manser, M. B. (2017). Exorcising Grice's ghost: An empirical approach to studying intentional communication in animals. *Biological Reviews*, *92*(3), 1427–1438. https://doi.org/10.1111/brv.12289

Veldhuizen, M. G., Farruggia, M. C., Gao, X., Nakamura, Y., Green, B. G., & Small, D. M. (2020). Identification of an Amygdala–Thalamic Circuit That Acts as a Central Gain Mechanism in Taste Perception. *Journal of Neuroscience*, *40*(26), 5207–5217. https://doi.org/10.1523/JNEUROSCI.2618-19.2020

Walther, C. C. (2021). *Technology, Social Change and Human Behavior: Influence for Impact*. Palgrave Macmillan Cham. https://doi.org/10.1007/978-3-030-70002-7

Wallace, C. S. (1999). Minimum Message Length and Kolmogorov Complexity. *The Computer Journal*, *42*(4), 270–283. https://doi.org/10.1093/comjnl/42.4.270

Wang, P., Hahm, C., & Hammer, P. (2022). A Model of Unified Perception and Cognition. *Frontiers in Artificial Intelligence*, *5*. https://www.frontiersin.org/articles/10.3389/frai.2022.806403

Wang, Y.-X., & Yin, B. (2023). A new understanding of the cognitive reappraisal technique: An extension based on the schema theory. *Frontiers in Behavioral Neuroscience*, *17*. https://www.frontiersin.org/articles/10.3389/fnbeh.2023.1174585

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., & Cacioppo, J. T. (2010). Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, *99*(3), 410–435. https://doi.org/10.1037/a0020240

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 201–233. https://doi.org/10.1037/xlm0000732

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An Internal Model for Sensorimotor Integration. *Science*, *269*(5232), 1880–1882. https://doi.org/10.1126/science.7569931

Wu, Y., Baker, C. L., Tenenbaum, J. B., & Schulz, L. E. (2018). Rational Inference of Beliefs and Desires From Emotional Expressions. *Cognitive Science*, *42*(3), 850–884. https://doi.org/10.1111/cogs.12548

Wunsch, A., Philippot, P., & Plaghki, L. (2003). Affective associative learning modifies the sensory perception of nociceptive stimuli without participant's awareness. *Pain*, *102*(1), 27–38. https://doi.org/10.1016/s0304-3959(02)00331-7

Yin, B., Wang, Y.-X., Fei, C.-Y., & Jiang, K. (2022). Metaverse as a possible tool for reshaping schema modes in treating personality disorders. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.1010971

Yongsatianchot, N., Thejll-Madsen, T., & Marsella, S. (2023). What's Next in Affective Modeling? Large Language Models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (pp. 1-7). IEEE. https://doi.org/10.1109/ACIIW60385.2023.10350806

Yu, W., Xu, H., Yuan, Z., & Wu, J. (2021). Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(12), 10790–10798. https://doi.org/10.1609/aaai.v35i12.17289

Zajonc, R. B. (1984). On the primacy of affect. *American Psychologist*, *39*(2), 117–123. https://doi.org/10.1037/0003-066X.39.2.117

Zhao, K., Chu, W.-S., & Zhang, H. (2016). Deep region and multi-label learning for facial action unit detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3391–3399. https://doi.org/10.1109/CVPR.2016.368

Zhao, Y., Xu, L., Huang, Z., Peng, K., Seligman, M., Li, E., & Yu, F. (2023). AI chatbot responds to emotional cuing [Preprint]. *In Review*. https://doi.org/10.21203/rs.3.rs-2928607/v1

Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., Wang, H., Pang, Y., Jiang, W., Zhang, J., Li, Z., Zhang, W., Li, Z., Liu, W., & Yuan, L. (2023). LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. *arXiv preprint arXiv:2310.01852*.

Note: Figure translations are in progress. See original paper for figures.