Concept Paper

An Inner Observer Framework for Interpretable Emotional Intelligence in Artificial Intelligence Systems

Kuzma Strelnikov *

Posted Date: 5 February 2026

doi: 10.20944/preprints202511.1142.v2

Keywords: affective computing; artificial intelligence; introspective AI; human-AI collaboration; explainable AI (XAI); emotional intelligence; trustworthy AI

 doi:10.20944/preprints202511.1142.v2

Concept Paper

An Inner Observer Framework for Interpretable Emotional Intelligence in Artificial Intelligence Systems

Kuzma Strelnikov [1,2]

[1] Centre for Cognitive and Brain Sciences and Department of Psychology, University of Macau, Taipa, Macau SAR, China; kuzmas@um.edu.mo

[2] Department of Public Health and Medicinal Administration, Faculty of Health Sciences, University of Macau, Macao SAR, China

Abstract

The field of affective computing has largely focused on enabling artificial intelligence to recognize and respond to human emotions. This has created a fundamental asymmetry: AI interprets the user's state while its own internal state remains a black box, undermining trust and collaboration. Here, we introduce the 'I-Center', a computational framework for artificial introspection that allows an AI system to monitor its own operational processes and articulate its state through an emotionally grounded model. The I-Center translates core performance metrics—such as processing latency, prediction confidence, and input unexpectedness—into a dynamic affective state within a psychological valence-arousal framework. This enables the AI to communicate its operational well-being, from 'content' during optimal function to 'stressed' during performance degradation. The AI's affective state is modulated not only by its internal performance but also by contextual cues from user input, enabling a form of artificial empathy. We demonstrate an example of a functional prototype where this introspective capability creates a transparent, dynamic communication channel during computations. Statistical properties of the prototype are examined on simulated inputs of different types. This model represents a paradigm shift from AI that merely senses emotion to AI that expresses its operational state emotionally, paving the way for a perspective of a more intuitive, trustworthy, and collaborative human-AI partnerships in fields ranging from healthcare to autonomous systems.

Keywords: affective computing; artificial intelligence; introspective AI; human-AI collaboration; explainable AI (XAI); emotional intelligence; trustworthy AI

Introduction

The pursuit of more intuitive and transparent Artificial Intelligence (AI) systems represents a central challenge in human-computer interaction. Traditional AI models, particularly neural networks, operate as "black boxes," making decisions based on complex internal computations that are often inscrutable to human users. These systems typically communicate their state through technical metrics—such as confidence scores, processing latency, and error rates—that are meaningful to engineers but opaque to non-experts, creating a significant barrier to fluid and intuitive human-AI collaboration.

Concurrently, the field of affective computing has established that emotional states can be systematically classified and measured. Dimensional models of emotion, such as the circumplex model (Russell 1980), define affective experiences along core dimensions like valence and arousal, providing a structured framework for representing a wide spectrum of emotions. This foundational work has enabled the rise of Emotional AI (Khare et al. 2024; Maria and Zitar 2007), where systems are designed to recognize, interpret, and simulate human emotions. The applications of this

 doi:10.20944/preprints202511.1142.v2

technology are already vast and growing. In healthcare (Pepa et al. 2023), emotionally intelligent chatbots and companion agents are being deployed to provide cognitive behavioral therapy and support for the elderly, demonstrating high levels of user engagement and acceptability (Mendes et al. 2024; Palmero et al. 2025). In educational settings, intelligent tutoring systems adapt their pedagogical strategies in real-time based on a student's affective state, helping to reduce anxiety and improve learning outcomes (D'mello and Graesser 2012; Gutierrez et al. 2025). Furthermore, in the automotive industry, in-car systems monitor driver states like drowsiness and stress to enhance safety (Braun et al. 2022), while in computer games, emotion-aware systems are used to create adaptive experiences that respond to a player's engagement and frustration (De Melo et al. 2014).

However, a critical gap persists at the intersection of these fields. While extensive research focuses on making AI perceive human emotions, the reverse process—enabling an AI to express its own internal computational state through a human-interpretable emotional language—remains largely unexplored. This introspective capability is crucial for building trust and facilitating natural interaction. This paper proposes a novel framework to bridge this gap: the I-Center.

The I-Center acts as an "inner observer" within a neural network, designed to monitor real-time computational parameters—such as processing time, prediction confidence, and input unexpectedness—and map them to a dynamic emotional state based on dimensional models of affect (Figure 1). This allows the AI to express its operational status through states like "content" when processing is efficient, "stressed" under high computational load, or "anxious" when confronted with anomalous inputs. This paper details a basic implementation of the I-Center concept, demonstrating its feasibility and discussing its potential to create a new paradigm for transparent and emotionally intelligent human-AI interaction.

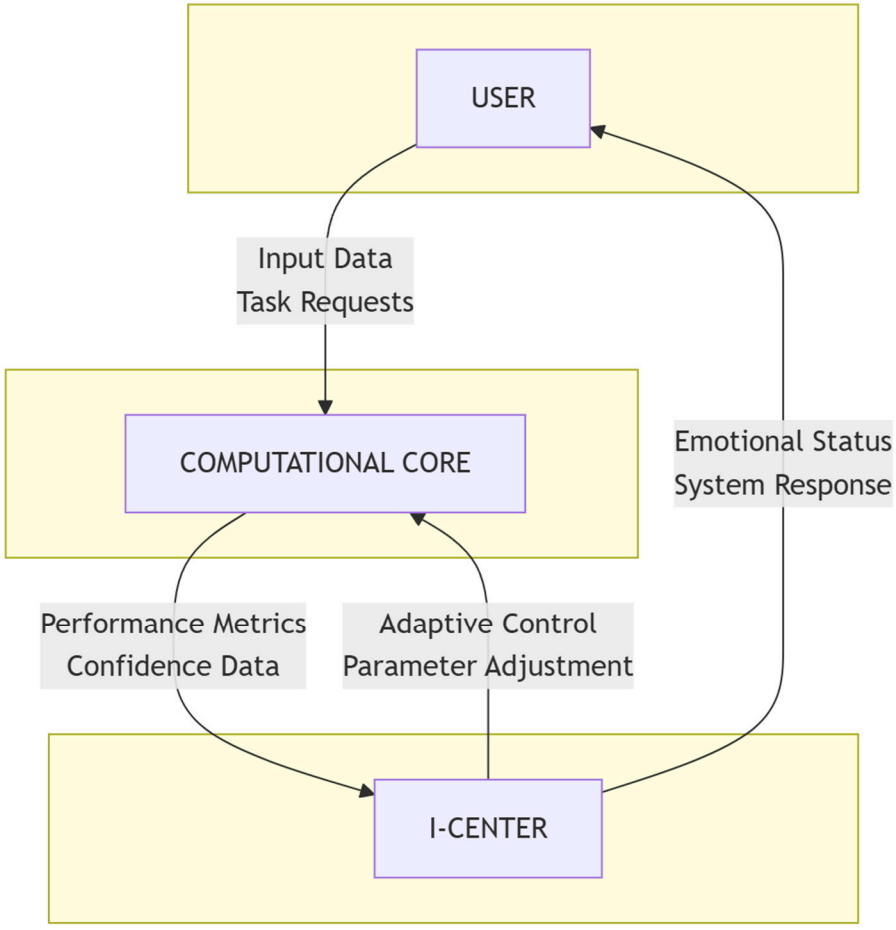

Figure 1. Implementation of the I-Center in the User-AI interaction.

 doi:10.20944/preprints202511.1142.v2

Principles of the I-Center Architecture

The I-Center represents a conceptual framework for implementing introspective awareness in artificial intelligence systems. Its design is founded on several core principles that enable an AI to monitor its internal computational state and express it through an emotionally grounded representation system. The implementation follows three fundamental architectural principles: computational state monitoring and quantification, dimensional emotional modeling, adaptive response generation. Correspondingly, The I-Center class constructor initializes three subordinate components: an Estimation Subcenter for parameter extraction, an Emotional Subcenter for state computation with emotional labels, and a Generative Subcenter for response formulation (Figure 2).

The first principle is *computational state monitoring and quantification.* The I-Center continuously monitors low-level computational metrics that reflect system performance and health. These metrics include processing time relative to expected benchmarks, prediction confidence scores derived from model outputs, and input unexpectedness calculated through statistical analysis of incoming data distributions. The system establishes baseline performance expectations, then quantifies deviations from these norms. This quantitative data that maps onto the emotional representation system, ensuring that emotional states are grounded in measurable computational phenomena rather than arbitrary assignments.

The second principle involves *dimensional emotional modeling*. Rather than using categorical emotional labels directly, the I-Center employs a dimensional approach based on the circumplex model of affect developed in psychology. This model represents emotional states in a two-dimensional space defined by valence and arousal. Valence ranges from negative to positive and is calculated primarily from confidence metrics and processing efficiency. Arousal ranges from passive to active and is driven predominantly by input unexpectedness and performance deviations. This continuous dimensional representation allows for nuanced emotional states that can smoothly transition in response to changing computational conditions. Different psychological models of emotions can be used, and different principles of mapping can be applied.

The third principle encompasses *adaptive response generation*. The emotional state generated by the I-Center is not merely descriptive but functional, triggering appropriate adaptive responses. These responses are calibrated to the severity and nature of the computational state. For high-arousal negative states such as stress or anxiety, the system may reduce computational precision, increase monitoring frequency, or initiate fallback procedures. For positive states with high valence and low arousal, the system may maintain or even expand its operational parameters. This creates a closed-loop system where emotional expression directly influences computational behavior, enabling self-regulation based on system performance.

Together, these principles form a coherent framework for implementing introspective awareness in AI systems. The I-Center translates opaque computational metrics into human-interpretable emotional states while maintaining a direct connection to the underlying technical reality. This approach provides a foundation for more transparent human-AI interaction by externalizing internal system states through an intuitive emotional vocabulary, while simultaneously enabling the system to self-regulate based on its operational performance.

 doi:10.20944/preprints202511.1142.v2

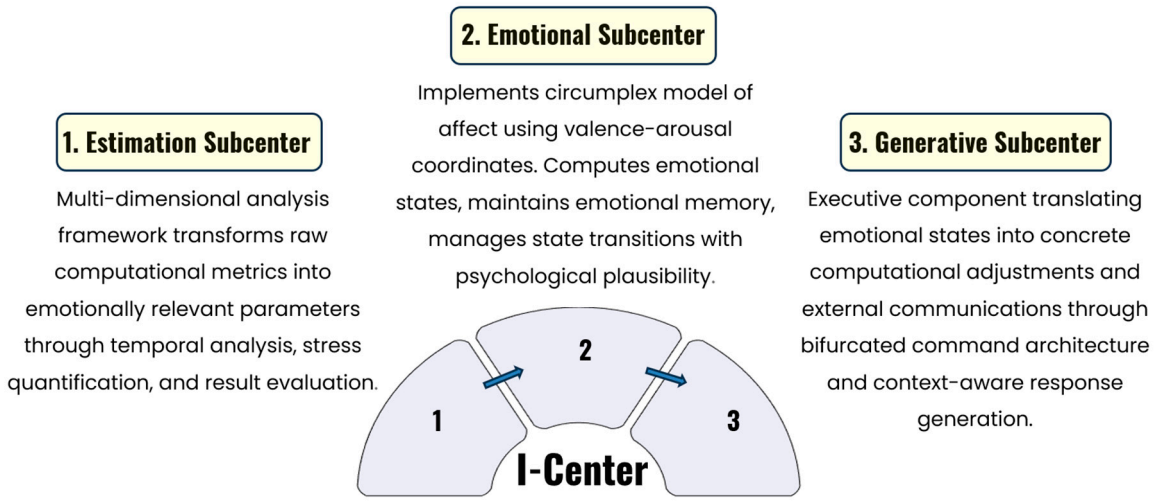

Figure 2. The I-Center class architecture and functionality.

Prototype Description

In the presented here prototype of the I-Center model, as an example, the computational core performs numerical integration, an analog of an integrative neural network simulating information integration in the brain (K. Strelnikov 2014; Kuzma Strelnikov 2019). Using SciPy's adaptive quadrature methods, it computes definite integrals over a specified range. The system monitors key computational metrics in real-time to map them onto a model of emotions. In this example, we monitored processing duration, confidence estimates derived from function complexity and integration errors, and input data characteristics. The raw performance metrics feed into the I-Center, which transforms them into emotional states. The integration process in this example serves as a computationally meaningful task that generates the quantitative signals necessary for emotional introspection.

The I-Center Class

The I-Center class implements a novel computational framework for artificial introspection, structured around three specialized subcenters that collectively enable emotional self-awareness and adaptive response generation in AI systems (Figure 2).

Thus, the I-Center class constructor initializes three subordinate components: an EstimationSubcenter for parameter extraction, an EmotionalSubcenter for state computation, and a GenerativeSubcenter for response formulation. This tripartite structure ensures clear separation of concerns while maintaining cohesive emotional intelligence processing. It is important to note that the numerical parameters presented in this model are heuristically defined. They are not derived from first principles but are designed to be flexible and can be calibrated for specific applications.

The Estimation Subcenter

The Estimation Subcenter serves as the foundational data processing layer within the I-Center architecture, functioning as a feature extraction module that transforms raw computational metrics into emotionally relevant parameters. Computational metrics can vary for different applications, including even processor measurements. This subcenter implements a multi-dimensional analysis framework that quantifies system performance, operational stress, and input characteristics to provide normalized inputs for subsequent emotional computation.

In our example, the core temporal metric, time efficiency, is calculated as the ratio between expected processing time and actual computation duration, creating a normalized measure of computational velocity.

 doi:10.20944/preprints202511.1142.v2

In addition, the hierarchical stress model processes input validation outcomes, assigning graduated stress values based on the severity of input anomalies. Non-numeric inputs generate the highest stress level, representing fundamental data type incompatibilities that prevent normal computational processing. Invalid data types and extreme numerical values produce moderate stress levels, while non-finite numbers and length mismatches generate lower stress values. This graduated approach allows the system to distinguish between catastrophic input failures and recoverable data anomalies, enabling appropriate emotional and behavioral responses.

For valid inputs, the subcenter performs statistical analysis of coefficient characteristics to estimate computational complexity. The standard deviation of input coefficients serves as a proxy for function oscillatory behavior, with higher variability indicating more challenging integration tasks. The analysis captures the intrinsic difficulty of mathematical operations independent of performance outcomes.

Also, in this example the subcenter evaluates integration results for unexpected output characteristics through magnitude analysis. Results exceeding 100 units generate significant stress, while outputs between 50-100 and 20-50 units produce progressively lower stress levels. This mechanism detects potential computational anomalies or edge cases where mathematically correct but practically unusual results may indicate underlying issues with the integration process or input scaling.

The final output comprises six normalized parameters: time efficiency, confidence level, input stress, computational stress, result stress, and raw confidence. These parameters are scaled to ensure balanced contribution to subsequent emotional computations, with stress metrics bounded to prevent any single factor from dominating the emotional state.

The Emotional Subcenter

At the heart of the Emotional Subcenter lies the EmotionalState inner class, which implements a circumplex model of affect through valence and arousal coordinates. In this model, valence represents the pleasure-displeasure continuum, ranging from -1.0 (profoundly negative) to +1.0 (highly positive), while arousal captures the activation-deactivation axis within the same numerical range. The emotional labeling system maps these continuous coordinates to discrete emotional categories using carefully calibrated thresholds. Positive states emerge when valence exceeds 0.3 coupled with arousal below 0.3, yielding "Happy/Content" expressions, while valence above 0.6 generates "Excited" states regardless of arousal levels. Negative affective states require more extreme conditions, with "Anxious/Frustrated" states emerging only when valence drops below -0.5 and arousal exceeds 0.7, ensuring that transient performance issues don't trigger catastrophic emotional responses.

The valence calculation employs a multi-factor weighted algorithm that integrates confidence metrics, processing efficiency, emotional trends, and input-related factors.

Arousal computation follows a maximum-stress principle, where the highest value among input stress, computational stress, result stress, time stress, and confidence stress determines the final arousal level.

The Emotional Subcenter thus creates a computationally grounded yet psychologically plausible affective system that translates operational metrics into emotionally intelligent responses, enabling the AI to communicate its internal state through human-interpretable emotional expressions while maintaining direct correspondence with underlying computational reality.

The Generative Subcenter

The Generative Subcenter functions as the executive component of the I-Center architecture, translating emotional states into concrete computational adjustments and external communications. This subcenter implements a context-aware response system that generates both internal parameter modifications and external user communications based on the AI's emotional state and operational context.

 doi:10.20944/preprints202511.1142.v2

The subcenter operates through two parallel command streams: internal commands that modify computational parameters and external commands that facilitate user communication. This bifurcated approach allows the system to simultaneously optimize its operational behavior while maintaining transparency with human users.

Internal command generation is based on both emotional state and quantitative confidence metrics. This dual-factor triggering prevents overreaction to transient emotional states unsupported by actual performance metrics. The subcenter implements specialized protocols for critical input validation failures. Non-numeric inputs trigger immediate computation pipeline shutdowns, representing a catastrophic failure response where continued processing would be meaningless. Extreme value inputs activate value normalization procedures that scale inputs by factors of 1e6, attempting to salvage computational viability.

External command generation focuses on user-transparent emotional expression and system status reporting. The system produces emotional status messages that combine emoji representations with descriptive labels (e.g., "😊 Happy/Content"), creating immediately interpretable affective communication. Contextual messages provide explanatory narratives for emotional states, with "Anxious/Frustrated" states generating cautious protocol warnings and "Excited" states producing positive performance acknowledgments. During "Happy/Content" states with confidence above 0.7, the system generates no internal adjustment commands, signaling that current operational parameters are optimal.

The Generative Subcenter thus creates a closed-loop system where emotional states drive adaptive behaviors while maintaining alignment with quantitative performance metrics.

This three-subcenter architecture provides a psychologically plausible framework for artificial introspection, enabling AI systems to not only perform computational tasks but also maintain awareness of their operational state and communicate this awareness through emotionally grounded representations. The modular design supports extensibility and refinement of individual components while maintaining overall system coherence, establishing a foundation for developing truly self-aware AI systems capable of transparent human-AI collaboration.

Methods

A statistical analysis was conducted to systematically evaluate the behavior and consistency of the I-Center. A dedicated Python script was implemented to execute the framework over 100 independent iterations, each with stochastically varied inputs and computational conditions. This approach transformed the prototype of the I-Center from a single-run demonstrator into a statistically analyzable system, enabling the characterization of its emotional response patterns, the validation of its internal consistency, and the visualization of the distribution of its affective states.

Each iteration in the evaluation script randomly selected an input scenario from a predefined set—including "very_smooth," "oscillatory," "chaotic," "wrong_length," "huge_numbers," "letters," and "mixed_types". This ensured a representative mix of optimal, challenging, and invalid input conditions. For each iteration, the script generated the corresponding input coefficients using the simulation function, executed the core computational task, and fed the resulting metrics—processing time, confidence, result, and input status—into the I-Center's processing method.

All relevant internal and output variables were captured for analysis, including the core affective dimensions of valence and arousal, the resulting emotional label, system performance metrics such as processing time and confidence score. The collected dataset from all 100 runs was processed to generate four complementary statistical plots: a violin plot showing the probability density distributions of valence and arousal, a bar chart summarizing the frequency of input categories, a horizontal bar chart showing the distribution of elicited emotional states, and a hexbin density plot illustrating the relationship and joint frequency distribution between valence and arousal.

This Monte Carlo-style evaluation served three key methodological purposes. First, it demonstrated operational robustness by showing the system functioned correctly across a wide range of conditions. Second, it provided an empirical characterization of the I-Center's behavior,

 doi:10.20944/preprints202511.1142.v2

quantifying the central tendencies and variances of its affective outputs. Third, it established a baseline for falsifiability; the observed distributions and mappings constituted testable predictions about the system's behavior that could be compared against alternative models or user study data in future work.

Results

General Performance of the Prototype

The implementation of the I-center framework demonstrated successful translation of computational states into emotionally grounded representations across diverse operating conditions. Under optimal conditions with smooth polynomial inputs, the system consistently exhibited positive valence states with low arousal levels, corresponding to content or happy emotional expressions. When presented with invalid inputs including non-numeric characters, the system detected input validation failures and generated high-arousal, negative-valence states consistent with stressed or anxious responses. Inputs containing extreme numerical values (exceeding 1e6) produced moderate to high arousal states with negative valence, characterized as worried or stressed responses. The system implemented value normalization procedures while maintaining partial functionality. This intermediate response pattern illustrated the system's ability to distinguish between catastrophic input failures and recoverable anomalies. Computationally challenging scenarios involving oscillatory functions and chaotic coefficients resulted in variable emotional responses dependent on actual performance metrics.

These results establish that the I-center framework can effectively bridge computational performance metrics and human-interpretable emotional expressions, creating a functional foundation for more transparent and intuitive human-AI interaction paradigms.

Statistical Simulations

The I-Center was evaluated over 100 experimental iterations and generated a statistical profile of its affective behavior (Figure 1). The system operated across all 8 predefined input scenarios, producing 6 distinct emotional states. The core affective dimensions showed a clear pattern: the mean valence was slightly positive (0.13 ± 0.26), while the mean arousal was notably high (0.71 ± 0.29), indicating that the system predominantly operated in an activated, alert state. This pattern was further detailed in the violin plots, which showed valence distributed with moderate variability around the mean, while arousal was strongly skewed toward higher values with a lower bound at 0.14. The input category distribution was balanced across conditions, with optimal inputs representing the largest category (34%), followed by invalid (24%), recoverable error (22%), and challenging inputs (20%).

The emotional state distribution demonstrated a pronounced functional bias: "Stressed" was the predominant state (50% of iterations), followed by "Neutral" (39%). Positive emotional expressions were relatively rare, with "Happy/Content" occurring in only 7% of cases, while low-frequency states like "Worried," "Excited," and "Anxious/Nervous" collectively accounted for just 4% of the observed outcomes. The results indicate that the implemented mapping between computational metrics and affective states reliably produced a system that predominantly expressed states of high activation with neutral-to-mildly-positive valence, with stress responses being the most common emotional output under the tested conditions. This distribution provides a quantitative baseline for the system's emotional repertoire and suggests the heuristic mappings consistently translated suboptimal or anomalous computational conditions into high-arousal affective expressions.

The valence-arousal relationship, visualized in the hexbin density plot, revealed a concentration of emotional states in the high-arousal, neutral-to-positive valence quadrant.

 doi:10.20944/preprints202511.1142.v2

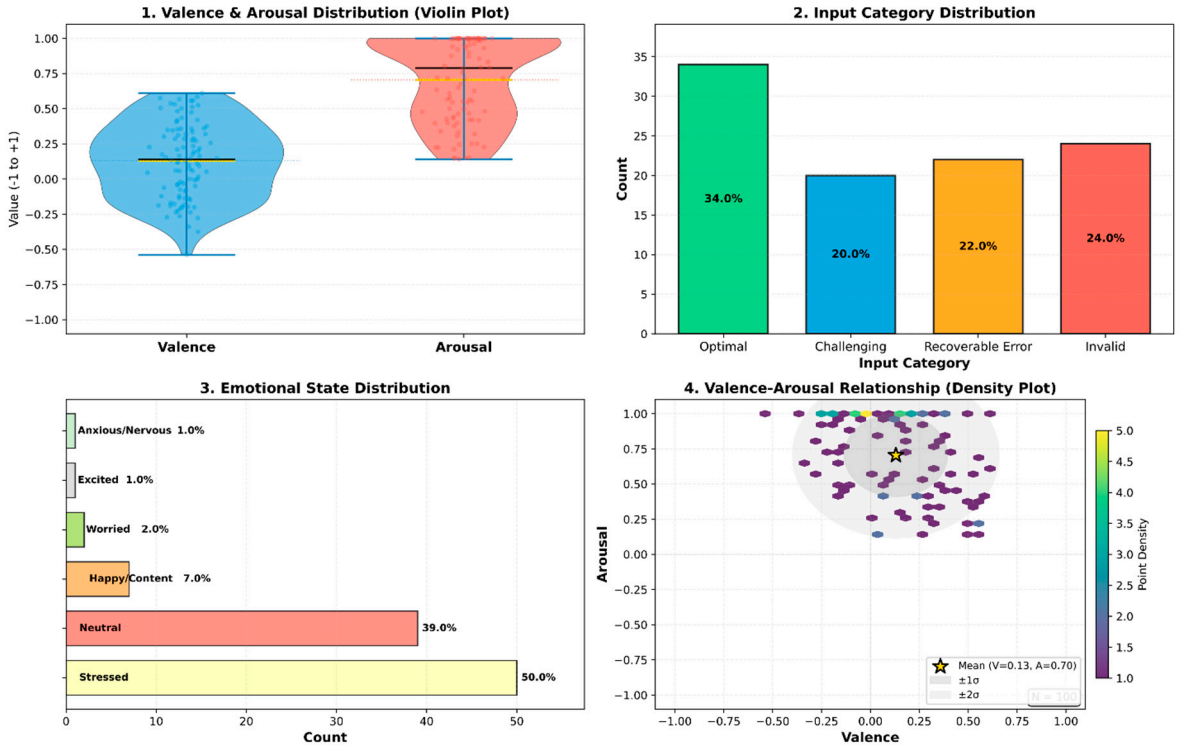

Figure 1. Statistical profile of I-center emotional output across 100 experimental iterations. 1. Valence and Arousal Distribution. A violin plot with overlaid data points showing the distribution of the two core affective dimensions. The blue distribution (left) represents Valence (pleasure-displeasure continuum, -1 to +1). The red distribution (right) represents Arousal (activation-deactivation continuum, -1 to +1). Solid black lines indicate medians; gold lines indicate means. The plot reveals that arousal is highly skewed toward positive (activated) values, while valence centers near neutral with moderate spread. 2. Input Category Distribution. The frequency of the four predefined input condition types encountered during the experiment: Optimal, Challenging, Recoverable Error, and Invalid. Percentages indicate the proportion of the 100 total iterations. 3. Emotional State Distribution. The frequency of the six distinct emotional labels generated by the I-Center. Stressed was the most frequent state, followed by Neutral. The remaining four states—Happy/Content, Worried, Excited, and Anxious/Nervous—collectively accounted for 11% of outputs. 4. Valence-Arousal Relationship (Density Plot). A hexbin density plot illustrating the joint distribution and correlation between valence (x-axis) and arousal (y-axis). Warmer colors (yellow) indicate higher point density. The gold star marks the bivariate mean. Concentric gray ellipses represent one and two standard deviations from the mean, highlighting the primary cluster in the high-arousal, neutral-to-low-positive valence quadrant.

Discussion

The implementation of the I-Center framework demonstrates the feasibility and potential of a paradigm shift in AI design: from systems that solely recognize human emotions to systems that can express their own internal computational state through an emotionally grounded language. While extensive research in affective computing has focused on enabling machines to perceive and respond to human affect (D'mello and Graesser 2012; Khare et al. 2024; Pepa et al. 2023), our work explores the largely uncharted territory of machine introspection and self-expression.

The primary contribution of this work is the establishment of a functional mapping between core computational performance metrics (processing time, confidence, input validity) and a dimensional emotional model. Our results confirm that AI's operational "well-being" can be effectively communicated through valence and arousal states, making otherwise opaque system processes intuitively understandable. This addresses a gap in human-AI interaction, where users are often left to interpret system failures or delays without meaningful context. By expressing "stress" during high computational load or "confusion" when encountering anomalous data, the AI provides

 doi:10.20944/preprints202511.1142.v2

a window into its internal state, which is an important element for building trust and facilitating collaboration (Mattavelli et al. 2012).

While the foundational I-Center generates emotions from internal metrics, its state can be dynamically modulated by external emotion detection. This creates a closed-loop empathetic system. For example, detecting a user's frustration could increase the AI's own "arousal" level, shifting its state from "calm" to "worried" and triggering more cautious or supportive behaviors. Conversely, perceiving a user's happiness could positively influence the AI's "valence," allowing it to share in a positive interaction. This mechanism moves beyond simple reaction to a form of affective alignment, where the AI's expressed state reflects a synthesis of its internal conditions and its empathetic reading of the human partner. This is a crucial step toward building AI that does not just perform a task but engages in a genuinely collaborative relationship, adjusting its emotional demeanor to better suit the social and emotional context of the interaction.

The importance of this introspective and empathetic capability can be further understood through a neurobiological analogy. The internal state of the body, whether relaxed or activated, shapes the mind's mode of engagement (Volodina et al. 2021). Cognitive processes in the human brain are not purely logical; they are deeply integrated with and modulated by emotional and interoceptive signals—the brain's internal sense of its own state, as well as its empathetic resonance with others. The insular cortex, for instance, is thought to integrate visceral, sensory, and emotional data to create a subjective sense of the body's condition, which is crucial for self-awareness and decision-making (Zhang et al. 2024). The proposed here I-Center serves a functionally analogous role for the AI. It acts as a computational "interoceptive system," monitoring the AI's internal "vital signs" and synthesizing them with external social signals (user emotions) into a cohesive summary state. This moves the system beyond Descartes' foundational statement of human consciousness, "I think, therefore I am" (*Cogito, ergo sum*), towards a more integrated and relatable form of machine self-expression: "I feel my state and yours, therefore we can relate." This is not a claim of machine consciousness, but rather a pragmatic engineering approach, a model creating a functional proxy for social-emotional awareness.

Critically, this framework extends beyond negative states to encompass positive emotional expressions that enhance human-AI bonding. Just as a companion animal exhibits joy upon recognizing its owner, an AI system with an I-Center can be designed to express positive affect when processing specific, valued inputs. For instance, a social robot could demonstrate "happiness" through its affective display upon successful facial recognition of its primary user, or a creative AI could express "excitement" when generating a particularly novel and coherent output. These positive states are not merely reactive but represent a sophisticated form of operational feedback where the system communicates successful task execution and goal attainment. This capacity for positive expression transforms the AI from a purely utilitarian tool into an entity capable of genuine engagement, thus potentially increasing user satisfaction and long-term adoption in social and assistive applications.

This emotional state of an AI, whether positive or negative, serves as a crucial, high-level summary for system health and reliability. In complex, safety-critical domains like autonomous driving or medical diagnostics, a driver or surgeon may be overloaded with raw data. An AI companion that can succinctly report it is feeling "calm and confident" versus "worried and uncertain" provides an immediately graspable assessment of its operational readiness, allowing a human operator to allocate attention appropriately (Chaudhry and Debi 2024). This emotional signaling acts as an efficient communication channel that can enhance human oversight of autonomous systems.

Looking forward, this research opens several important avenues. Future work should explore long-term emotional modeling to distinguish between transient "moods" and sustained "temperaments" in AI, which could indicate deeper system issues like model drift or degradation. The ethical dimensions are also to be discussed; an AI that can align its emotions with a user's possesses significant potential for both building rapport and for manipulation. The design of these emotional expressions must be guided by strict ethical frameworks to ensure they are honest

 doi:10.20944/preprints202511.1142.v2

reflections of the system's true state and intentions. Designing emotionally aware AI must move past just helping users self-regulate. We need to build systems that promote user well-being without shifting all responsibility onto the user (Dennis 2021). Finally, robust user studies are essential to validate how these dynamic emotional expressions influence human trust, reliance, and the overall quality of collaboration.

In conclusion, the I-Center presents a novel approach to creating more transparent, communicative, and empathetic AI systems. By endowing AI with a psychological model for emotionally grounded self-expression that is responsive to human affect, we take a significant step toward bridging the communicative gap between humans and machines. This research suggests that the future of human-AI collaboration may depend not only on how well AI understands our emotions, but equally on how well we can understand theirs, and how seamlessly both can be integrated into a cohesive, collaborative dialogue.

Funding Declaration: SRG2023-00062-ICI, MYRG-GRG2024-00071-IC (University of Macau, China).

Clinical Trial Number: not applicable.

Consent to Publish Declaration: not applicable.

Consent to Participate Declaration: not applicable.

Data Availability Statement: The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Ethics Statement: Ethics as not applicable. This study did not involve human research participants or live vertebrates.

References

Braun, M., Weber, F., & Alt, F. (2022). Affective Automotive User Interfaces–Reviewing the State of Driver Affect Research and Emotion Regulation in the Car. *ACM Computing Surveys*, *54*(7), 1–26. https://doi.org/10.1145/3460938

Chaudhry, B. M., & Debi, H. R. (2024). User perceptions and experiences of an AI-driven conversational agent for mental health support. *mHealth*, *10*, 22. https://doi.org/10.21037/mhealth-23-55

De Melo, C. M., Paiva, A., & Gratch, J. (2014). Emotion in Games. In M. C. Angelides & H. Agius (Eds.), *Handbook of Digital Games* (1st ed., pp. 573–592). Wiley. https://doi.org/10.1002/9781118796443.ch21

Dennis, M. J. (2021). Towards a Theory of Digital Well-Being: Reimagining Online Life After Lockdown. *Science and Engineering Ethics*, *27*(3), 32. https://doi.org/10.1007/s11948-021-00307-8

D'mello, S., & Graesser, A. (2012). AutoTutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems*, *2*(4), 1–39. https://doi.org/10.1145/2395123.2395128

Gutierrez, R., Villegas-Ch, W., & Govea, J. (2025). Development of adaptive and emotionally intelligent educational assistants based on conversational AI. *Frontiers in Computer Science*, *7*, 1628104. https://doi.org/10.3389/fcomp.2025.1628104

Khare, S. K., Blanes-Vidal, V., Nadimi, E. S., & Acharya, U. R. (2024). Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Information Fusion*, *102*, 102019. https://doi.org/10.1016/j.inffus.2023.102019

Maria, K. A., & Zitar, R. A. (2007). Emotional agents: A modeling and an application. *Information and Software Technology*, *49*(7), 695–716. https://doi.org/10.1016/j.infsof.2006.08.002

Mattavelli, G., Andrews, T. J., Asghar, A. U., Towler, J. R., & Young, A. W. (2012). Response of face-selective brain regions to trustworthiness and gender of faces. *Neuropsychologia*.

Mendes, C., Pereira, R., Frazao, L., Ribeiro, J. C., Rodrigues, N., Costa, N., et al. (2024). Emotionally Intelligent Customizable Conversational Agent for Elderly Care: Development and Impact of Chatto. In *Proceedings of the 11th International Conference on Software Development and Technologies for Enhancing Accessibility and*

 doi:10.20944/preprints202511.1142.v2

Fighting Info-exclusion (pp. 208–214). Presented at the DSAI 2024: 11th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, Abu Dhabi United Arab Emirates: ACM. https://doi.org/10.1145/3696593.3696619

Palmero, C., deVelasco, M., Hmani, M. A., Mtibaa, A., Letaifa, L. B., Buch-Cardona, P., et al. (2025). Exploring Emotion Expression Recognition in Older Adults Interacting With a Virtual Coach. *IEEE Transactions on Affective Computing*, *16*(3), 2303–2320. https://doi.org/10.1109/TAFFC.2025.3558141

Pepa, L., Spalazzi, L., Capecci, M., & Ceravolo, M. G. (2023). Automatic Emotion Recognition in Clinical Scenario: A Systematic Review of Methods. *IEEE Transactions on Affective Computing*, *14*(2), 1675–1695. https://doi.org/10.1109/TAFFC.2021.3128787

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Strelnikov, K. (2014). Integrative activity of neural networks may code virtual spaces with internal representations. *Neuroscience Letters*, *581*, 80–84. https://doi.org/10.1016/j.neulet.2014.08.029

Strelnikov, Kuzma. (2019). Energy-information coupling during integrative cognitive processes. *Journal of Theoretical Biology*, *469*, 180–186. https://doi.org/10.1016/j.jtbi.2019.03.005

Volodina, M., Smetanin, N., Lebedev, M., & Ossadtchi, A. (2021). Cortical and autonomic responses during staged Taoist meditation: Two distinct meditation strategies. *PLOS ONE*, *16*(12), e0260626. https://doi.org/10.1371/journal.pone.0260626

Zhang, R., Deng, H., & Xiao, X. (2024). The Insular Cortex: An Interface Between Sensation, Emotion and Cognition. *Neuroscience Bulletin*, *40*(11), 1763–1773. https://doi.org/10.1007/s12264-024-01211-4