Concept Paper

Internal Emotional Intelligence in AI Systems: An I-Center Framework for Human-Interpretable System States

Kuzma Strelnikov *

Posted Date: 17 November 2025

doi: 10.20944/preprints202511.1142.v1

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

 doi:10.20944/preprints202511.1142.v1

Concept Paper

Internal Emotional Intelligence in AI Systems: An I-Center Framework for Human-Interpretable System States

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. Crucially, our model is bidirectional; 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 a functional implementation where this introspective capability creates a transparent, dynamic communication channel. This work represents a paradigm shift from AI that merely senses emotion to AI that expresses its operational state emotionally, paving the way for 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 [1], 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 [2,3], where systems are designed to recognize, interpret, and simulate human emotions. The applications of this technology are already vast and growing. In healthcare [4], emotionally intelligent chatbots and companion agents are being deployed to provide

 doi:10.20944/preprints202511.1142.v1

cognitive behavioral therapy and support for the elderly, demonstrating high levels of user engagement and acceptability [5,6]. 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 [7,8]. Furthermore, in the automotive industry, in-car systems monitor driver states like drowsiness and stress to enhance safety [9], while in computer games, emotion-aware systems are used to create adaptive experiences that respond to a player's engagement and frustration [10].

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.

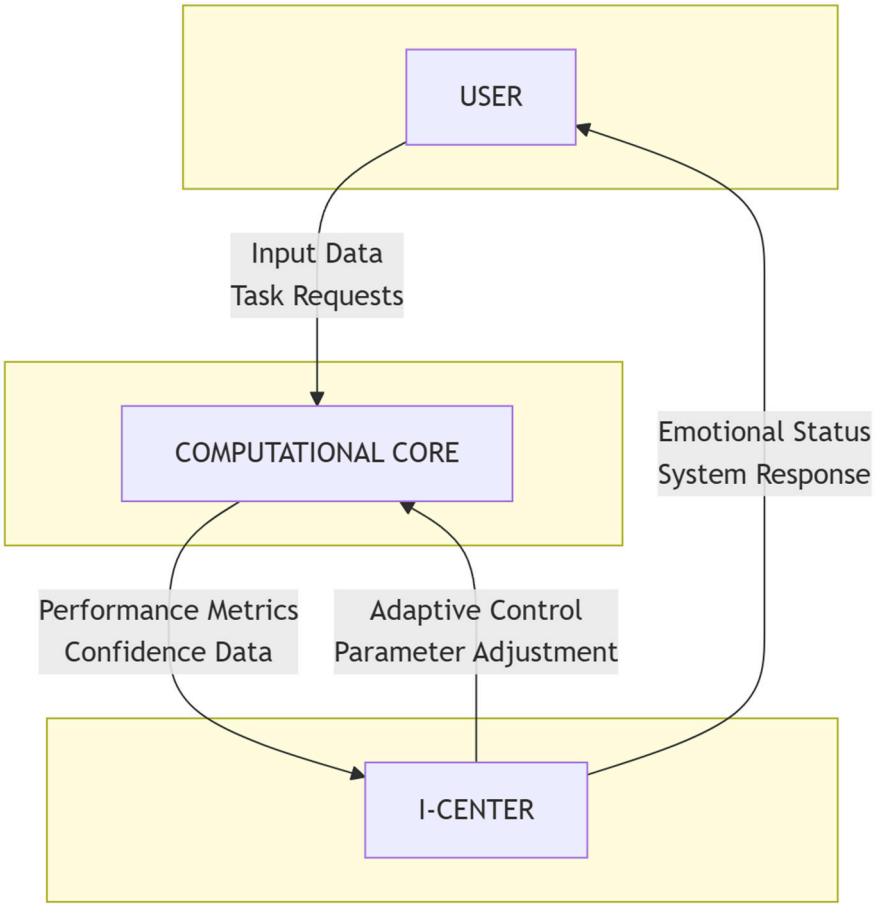


Figure 1. Implementation of the I-Center in the User-AI interaction.

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

 doi:10.20944/preprints202511.1142.v1

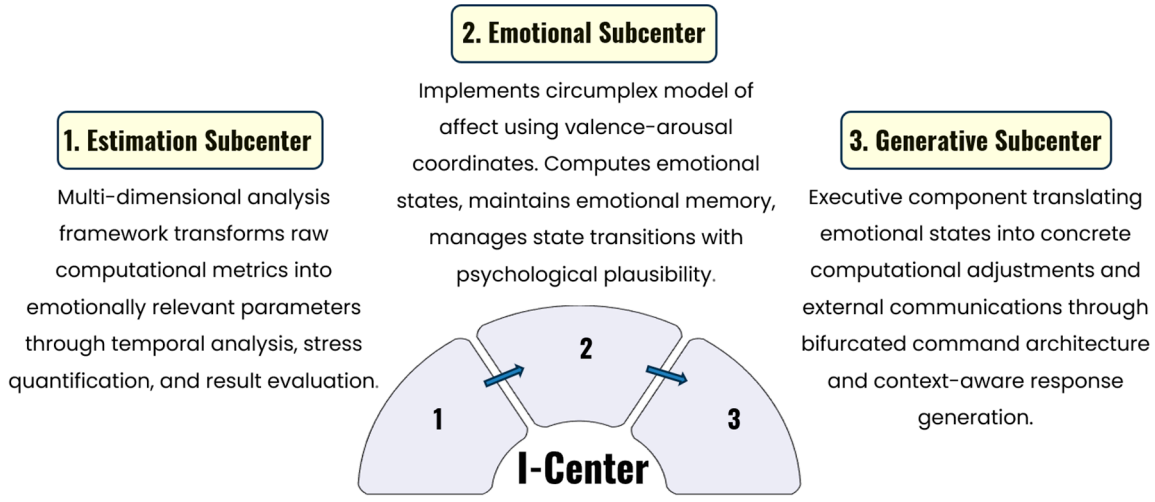
monitor its internal computational state and express it through an emotionally grounded representation system. The implementation follows three fundamental architectural principles: computational state monitoring and quantification, dimensional emotional modeling, adaptive response generation. Correspondingly, 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.v1

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

Methods

The computational core performs numerical integration, an analog of an integrative neural network simulating information integration in the brain [11,12]. Using SciPy's adaptive quadrature methods, it computes definite integrals over a specified range. The system monitors key computational metrics in real-time: processing duration, confidence estimates derived from function complexity and integration errors, and input data characteristics. These raw performance metrics feed into the I-Center, which transforms them into emotional states. The integration process serves as a computationally meaningful task that generates the quantitative signals necessary for emotional introspection, creating a direct link between mathematical operations and affective representation.

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 sophisticated feature extraction module that transforms raw computational metrics into emotionally relevant parameters. 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.

Temporal Performance Analysis

The subcenter maintains a sliding window history of computational performance metrics, including processing times, confidence scores, and input validation status. This historical context enables the system to distinguish between transient anomalies and persistent performance patterns. The core temporal metric, time efficiency, is calculated as the ratio between expected processing time (estimate_expected_processing_time function) and actual computation duration, creating a normalized measure of computational velocity that ranges from significantly suboptimal (values < 1.0) to highly efficient (values > 1.0). Confidence levels are aggregated through a moving average of recent confidence scores, providing a smoothed representation of the system's self-assessment reliability over time.

Input Stress Quantification

A 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 (>0.75), representing fundamental data type incompatibilities that prevent normal computational processing. Invalid data types and extreme numerical values produce moderate stress levels (e.g., 0.6), while non-finite numbers and length mismatches generate lower stress values (0.5 and 0.4). This graduated approach allows the system to distinguish between catastrophic input failures and recoverable data anomalies, enabling appropriate emotional and behavioral responses.

 doi:10.20944/preprints202511.1142.v1

Computational Complexity Assessment

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. This computational stress metric is bounded between 0.0 and 0.6, ensuring that even highly complex functions do not overwhelm the emotional computation system. The analysis captures the intrinsic difficulty of mathematical operations independent of performance outcomes.

Result Magnitude Evaluation

The subcenter evaluates integration results for unexpected output characteristics through magnitude analysis. Results exceeding 100 units generate significant stress (0.7), while outputs between 50-100 and 20-50 units produce progressively lower stress levels (0.4 and 0.2 respectively). 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.

Parameter Integration and Normalization

The final output comprises six normalized parameters: time efficiency, confidence level, input stress, computational stress, result stress, and raw confidence. These parameters are carefully scaled to ensure balanced contribution to subsequent emotional computations, with stress metrics bounded to prevent any single factor from dominating the emotional state. The parameter set provides a comprehensive snapshot of system operational health, capturing both immediate computational circumstances and emerging performance trends.

The Emotional Subcenter

Dimensional Emotion Framework

At the heart of the Emotional Subcenter lies the EmotionalState inner class, which implements a circumplex model of affect through valence and arousal coordinates. 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.

Valence Computation Algorithm

The valence calculation employs a multi-factor weighted algorithm that integrates confidence metrics, processing efficiency, emotional trends, and input-related factors. Confidence valence contributes 50% of the final valence value, calculated as twice the deviation from the 0.5 confidence baseline, ensuring that high confidence scores (above 0.7) strongly drive positive valence while low confidence (below 0.3) generates negative valence. Processing efficiency contributes 30% through a hyperbolic tangent transformation of time efficiency ratios, creating smooth valence transitions across performance levels. Emotional trends account for 10% of valence, capturing momentum in affective states by analyzing recent valence history, while input-related penalties contribute the final 10%, applying moderate valence reductions based on input stress levels. This balanced weighting ensures that confidence remains the primary valence driver while incorporating contextual performance factors.

 doi:10.20944/preprints202511.1142.v1

Arousal Regulation Mechanism

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. Time stress activates progressively, with processing delays exceeding 1.0 seconds generating 0.8 arousal, while milder delays (0.6-1.0 seconds) produce 0.4 arousal. Confidence stress follows a tiered approach, with confidence below 0.2 triggering high arousal (0.8), 0.2-0.4 generating moderate arousal (0.4), and 0.4-0.6 producing mild arousal (0.1). The system incorporates emotional noise through random variations between -0.1 and +0.1, introducing naturalistic variability while preventing erratic emotional oscillations.

Emotional Memory and Trend Analysis

The subcenter maintains an emotional memory buffer that stores recent valence-arousal pairs, enabling trend analysis that captures affective momentum. The trend calculation examines valence progression over the most recent emotional states, amplifying trends by a factor of 1.5 to emphasize developing emotional patterns while constraining values within the -1.0 to +1.0 range. This historical context prevents emotional instability while allowing the system to recognize and respond to persistent performance trends, whether positive or negative.

State Transition Management

Emotional state transitions incorporate smoothing mechanisms that prevent abrupt affective changes. The current implementation emphasizes emotional stability while remaining responsive to significant computational events. The threshold structure ensures that positive states are readily accessible during normal operation, while negative states require sustained or severe performance issues to activate. This design reflects a psychological principle where systems, like humans, maintain generally positive operational states unless confronted with substantial challenges or failures.

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.

Bifurcated Command Architecture

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 commands directly influence computational processes such as confidence thresholds, precision levels, and processing strategies, while external commands generate status messages, warnings, and performance metrics for user consumption.

Context-Aware Internal Command Generation

Internal command generation follows a sophisticated decision tree that considers both emotional state and quantitative confidence metrics. The system employs conditional logic where emotional labels trigger specific responses only when accompanied by supporting confidence evidence. For

 doi:10.20944/preprints202511.1142.v1

instance, "Anxious/Frustrated" states generate precision reduction commands and confidence threshold increases to 0.85, but only when raw confidence scores fall below 0.3. Similarly, "Stressed" states trigger moderate threshold increases to 0.8 when confidence drops below 0.4, while "Worried" states produce milder adjustments to 0.75 thresholds with confidence below 0.5. This dual-factor triggering prevents overreaction to transient emotional states unsupported by actual performance metrics.

Emergency Response Protocols

The subcenter implements specialized emergency 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 from otherwise unusable data. These emergency responses demonstrate the system's capacity for graded failure management, distinguishing between complete computational impossibilities and salvageable but challenging scenarios.

External Communication Strategy

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. The communication system employs urgency grading, with emergency states triggering high-urgency notifications while optimal operations generate low-urgency positive feedback.

Performance Metric Integration

All external communications include integrated performance metrics displaying current confidence levels and processing efficiency ratios. This ensures that emotional expressions remain grounded in quantitative reality, preventing potential misinterpretation of affective states. The metric display provides users with concrete data to contextualize emotional expressions, maintaining the system's credibility and transparency.

Response Calibration and Restraint

A key feature of the Generative Subcenter is its implementation of response calibration, where the absence of commands represents a positive outcome. During "Happy/Content" states with confidence above 0.7, the system generates no internal adjustment commands, signaling that current operational parameters are optimal. This restraint mechanism prevents unnecessary system modifications during peak performance periods, reducing computational overhead and maintaining stability.

The Generative Subcenter thus creates a sophisticated closed-loop system where emotional states drive adaptive behaviors while maintaining alignment with quantitative performance metrics. This ensures that the AI's emotional expressions translate into functionally appropriate responses, whether through internal optimization or external communication, establishing a coherent relationship between affective experience and operational reality.

Integrated Processing Pipeline

The I-Center's `process_iteration` method orchestrates the complete emotional computation cycle through sequential subcenter activation. The estimation phase transforms raw metrics into emotional parameters, the emotional phase computes the current affective state, and the generative

 doi:10.20944/preprints202511.1142.v1

phase produces both internal adjustments and external communications. This pipeline creates a closed-loop system where emotional states directly influence computational behavior while simultaneously enabling transparent communication of system status.

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.

Results

The implementation of the I-center framework demonstrated successful translation of computational states into emotionally grounded representations across diverse operating conditions. System performance and emotional responses were evaluated through a series of controlled computational scenarios designed to simulate both normal and edge-case 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. These states emerged when processing times remained within expected parameters and confidence scores exceeded 0.8. The system maintained stable operation with minimal adaptive adjustments, indicating recognition of efficient computational performance.

When presented with invalid inputs including non-numeric characters, the system immediately detected input validation failures and generated high-arousal, negative-valence states consistent with stressed or anxious responses. Processing times reduced significantly as the system implemented rapid rejection protocols, while confidence scores dropped to 0.02-0.05, reflecting appropriate uncertainty in handling fundamentally incompatible input types. The emotional response triggered emergency shutdown procedures, demonstrating the framework's capacity to escalate response severity appropriately.

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, with confidence scores averaging 0.1-0.3. 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. Extended processing times combined with moderate confidence scores (0.4-0.6) generated worried states, while more severe performance degradation triggered anxious or frustrated responses. The system appropriately adjusted confidence thresholds and computational precision in response to these emotional states.

The emotional reset mechanism functioned as designed, ensuring each computational iteration obtained its own state. There was no emotional carry-over between independent processing tasks while allowing appropriate emotional development within single computations. The system demonstrated consistent emotional trajectories based solely on current computational metrics without influence from prior emotional history (although this possibility could also be useful in some scenarios).

Across all test scenarios, the valence-arousal model effectively captured the multidimensional nature of system performance, with valence strongly influenced by confidence scores and arousal closely associated with processing time deviations and input unexpectedness. The discrete emotional labels provided intuitive categorization of system states while maintaining connection to the underlying continuous emotional dimensions.

The adaptive response system successfully translated emotional states into context-appropriate computational adjustments. Stress states triggered precision reduction and increased monitoring,

 doi:10.20944/preprints202511.1142.v1

recovery procedures activated during persistent poor performance, and optimal states permitted exploration of more complex computations. This demonstrated the closed-loop functionality of the emotional response system in regulating computational behavior.

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.

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 [2,4,7], 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 an AI's operational "well-being" can be effectively communicated through valence and arousal states, making otherwise opaque system processes intuitively understandable. This addresses a critical 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 a transparent window into its internal state, which is a foundational element for building trust and facilitating collaboration [13].

A pivotal advancement of our model is its capacity for **affective bidirectionality**. 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. In the human brain, cognitive processes 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 [14]. Our 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"

 doi:10.20944/preprints202511.1142.v1

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 [15]. 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 reflections of the system's true state and intentions. 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: Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

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

References

1. Russell JA. A circumplex model of affect. J Pers Soc Psychol. 1980;39(6):1161–78. https://doi.org/10.1037/h0077714
2. Khare SK, Blanes-Vidal V, Nadimi ES, Acharya UR. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. Inf Fusion. 2024;102:102019. https://doi.org/10.1016/j.inffus.2023.102019
3. Maria KA, Zitar RA. Emotional agents: A modeling and an application. Inf Softw Technol. 2007;49(7):695–716. https://doi.org/10.1016/j.infsof.2006.08.002

 doi:10.20944/preprints202511.1142.v1

4. Pepa L, Spalazzi L, Capecci M, Ceravolo MG. Automatic Emotion Recognition in Clinical Scenario: A Systematic Review of Methods. IEEE Trans Affect Comput. 2023;14(2):1675–95. https://doi.org/10.1109/TAFFC.2021.3128787
5. Mendes C, Pereira R, Frazao L, Ribeiro JC, Rodrigues N, Costa N, et al. Emotionally Intelligent Customizable Conversational Agent for Elderly Care: Development and Impact of Chatto. Proc 11th Int Conf Softw Dev Technol Enhancing Access Fight Info-Exclusion. Abu Dhabi United Arab Emirates: ACM; 2024. pp. 208–14. https://doi.org/10.1145/3696593.3696619
6. Palmero C, deVelasco M, Hmani MA, Mtibaa A, Letaifa LB, Buch-Cardona P, et al. Exploring Emotion Expression Recognition in Older Adults Interacting With a Virtual Coach. IEEE Trans Affect Comput. 2025;16(3):2303–20. https://doi.org/10.1109/TAFFC.2025.3558141
7. D'mello S, Graesser A. AutoTutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Trans Interact Intell Syst. 2012;2(4):1–39. https://doi.org/10.1145/2395123.2395128
8. Gutierrez R, Villegas-Ch W, Govea J. Development of adaptive and emotionally intelligent educational assistants based on conversational AI. Front Comput Sci. 2025;7:1628104. https://doi.org/10.3389/fcomp.2025.1628104
9. Braun M, Weber F, Alt F. Affective Automotive User Interfaces–Reviewing the State of Driver Affect Research and Emotion Regulation in the Car. ACM Comput Surv. 2022;54(7):1–26. https://doi.org/10.1145/3460938
10. De Melo CM, Paiva A, Gratch J. Emotion in Games. In: Angelides MC, Agius H, editors. Handb Digit Games. 1st ed. Wiley; 2014. pp. 573–92. https://doi.org/10.1002/9781118796443.ch21
11. Strelnikov K. Energy-information coupling during integrative cognitive processes. J Theor Biol. 2019;469:180–6. https://doi.org/10.1016/j.jtbi.2019.03.005
12. Strelnikov K. Integrative activity of neural networks may code virtual spaces with internal representations. Neurosci Lett. 2014;581:80–4. https://doi.org/10.1016/j.neulet.2014.08.029
13. Mattavelli G, Andrews TJ, Asghar AU, Towler JR, Young AW. Response of face-selective brain regions to trustworthiness and gender of faces. Neuropsychologia. 2012;
14. Zhang R, Deng H, Xiao X. The Insular Cortex: An Interface Between Sensation, Emotion and Cognition. Neurosci Bull. 2024;40(11):1763–73. https://doi.org/10.1007/s12264-024-01211-4
15. Chaudhry BM, Debi HR. User perceptions and experiences of an AI-driven conversational agent for mental health support. mHealth. 2024;10:22. https://doi.org/10.21037/mhealth-23-55