PUBLIC SUBMISSION

Received: May 29, 2025
Tracking No. mb9-se46-krc0
Comments Due: May 28, 2025 **Submission Type:** API

Docket: NSF-2025-OGC-0001
NITRD_FRDOC_0001

Comment On: NSF-2025-OGC-0001-0001
Request for Information: Development of a 2025 National Artificial Intelligence Research and Development Strategic Plan

Document: NSF-2025-OGC-0001-DRAFT-0243
Comment on FR Doc # 2025-07332

Submitter Information

Government Agency Type: State
Government Agency: University of California, San Diego

General Comment

AE4AI: Artificial Emotions for Artificial Intelligence

Biological Emotions (BE) are there for a reason. They are crucial aspects of evolution, playing a vital roles in cognitive processing under constraints and social cooperation. BE are needed and accompany behaviors in complex, uncertain environments, as well as in interpersonal interactions that are essential for survival. Emotions support Intelligence, survival and adaptation. A new line of research is needed to emphasize both the intrapsychic and interpersonal functions of emotions (in biological and artificial systems) rather than their experience or expression in/by agents. Not only do many (but not all) robots and AI systems need emotions, they should have them. It is time for Artificial Emotions (AEs).

See attached file(s)
..

Attachments

AE4AI-V4-Final

AE4AI: Artificial Emotions for Artificial Intelligence

Jean-Marc Fellous[1] and Eva Hudlicka[2]

1. University of California, San Diego, La Jolla, CA, USA
2. Psychometrix Associates & Therapy 21st, Amherst, MA, USA

Biological Emotions (BE) are there for a reason. They are crucial aspects of evolution, playing a vital roles in cognitive processing under constraints and social cooperation. BE are needed and accompany behaviors in complex, uncertain environments, as well as in interpersonal interactions that are essential for survival. Emotions support Intelligence, survival and adaptation. A new line of research is needed to emphasize both the intrapsychic and interpersonal **functions of emotions** (in biological and artificial systems) rather than their experience or expression in/by agents. **Not only do many (but not all) robots and AI systems need emotions, they should have them. It is time for Artificial Emotions (AEs).**

- Artificial Emotions for Ethical AI. One of the major obstacles to AI adoption is its potentially unethical or inappropriate outputs and lack of self-accountability. Ethical 'filters' (i.e., analyzing the outputs after they are generated) are only a limited first order way to manage this shortcoming. In addition, AI systems do not readily adapt to changes in the interaction context, typically rely on ad hoc strategies and can easily be fooled and manipulated. While AEs are not the full answer to these difficult issues they do **offer a promising path towards directing/modulating reasoning in an 'ethical' and robust fashion**. An artificial emotional system can support an implementation of empathy, which is the result of complex emotional processing, and thereby be a key contributor to 'ethical' behavior.

- Artificial Emotions for Specialized and Individualized Artificial Intelligence. Current AI is about averages, extracted from vast amounts of training datasets and represented in terms of static, statistical features. However, intelligence is not just knowledge. It also includes the manner in which the knowledge is used (retrieved, processed, modified or discarded). Intelligence, especially the type of intelligence required to generate original (i.e. eccentric, not average) solutions to common problems, relies on individual differences and the associated idiosyncratic problem-solving approaches. Emotions represent one way to express/implement such individual differences and the associated creativity. **The need for AEs will be more prominent as the AI research focuses away from General Artificial Intelligence (GAI) towards Specialized Artificial Intelligence (SAI).**
  Future SAI will feature systems with domain specific knowledge (e.g. Chess playing, Medical assistant), but accompanied with individual 'styles' and 'preferences' in solving particular problems. Part of this individuality will be implemented, we argue, as Artificial Emotions learned and tailored to the specific domain under consideration (e.g. fear of a bad chess position, fear of a life-long debilitating medical outcome).

- Beyond Recognition and Expression of Emotions: Affective computing, the subfield of AI addressing emotions, is almost exclusively focused on the communicative functions of emotions: their recognition and expression. Modern AI systems can detect and classify many human emotions using multimodal data, and robots and virtual agents can be made to simulate human modes of expressing emotions very effectively. This has definite advantages for human-agent interactions. However, very few artificial systems actually **use AEs to improve their functions**. The development of AEs in AI systems and agents is in its infancy. The development of intelligent, adaptive, transparent

and explainable AI systems needs to emphasize synergistically the development of AEs and the type of reasoning/decision making that includes emotional processing.

- Emotions are Emergent Multi-dimensional Constructs, not Categories. The major problem of most current approaches to understanding emotions (biological or artificial) is that researchers start by assessing and explaining the expression of emotions, walking their way backwards to their cause. This is a similar approach to understanding a mental health disorder by analyzing its symptoms. Problems are: 1) two individuals can have similar biological emotional expressions due to very different underlying neural processes/reasons, 2) an expressed emotion can be the result of another (non-expressed) emotion: emotions may co-occur and/or be masked, 3) classification of emotions by their expressions is somewhat arbitrary and culture/individual dependent. Emotions should be analyzed in a multi-dimensional framework. It is important to avoid using a classification (DSM-like) approach to understanding emotions and use a developmental/dynamical/transdiagnostic, and ultimately architecture-based, approach instead.

  A new approach is needed that is focused on the way behaviors and neural processes are changing and studying the cause/effect of these changes (some of which will be 'emotional'). **Artificial Emotions are/should be intrinsically part of Artificial Intelligence.** Computational systems are natively multi-dimensional and offer therefore a perfect environment to implement, understand and use AEs.

  The labeling of an emotion as 'fear' or 'happiness' is a convenient, but grossly oversimplified characterization of the underlying phenomenon. Emotions are emergent phenomena that can only be described in very high dimensional spaces. The dimensions are to this day unknown, but may in part correlate with some neural sub-circuits (e.g. PFC-amygdala interaction strength) or neural mechanisms (e.g. 5HT neuromodulation in cortex). These dimensions, in and of themselves, do not constitute emotions per se (i.e. there are no basic emotions, no dedicated emotion brain areas, but there are basic emotional components/processes), and may be different from person to person. The manifold on which the emotions reside has different shapes in different animals, and among different individuals of the same species. It also changes with age, and with mental health status. Artificial emotions should follow the same principles. **Artificial Emotions are not computable**: There are no general algorithms that can take inputs (as complex as they may be) and determine the emotional state. Only approximations/convincing 'fakes' can be made by such methods.

- Artificial Emotions are not Necessarily Human or Animal Emotions. Fear in a rat is not fear in a human. Both may have similarities (in expression or experience) but they are not the same. There are human emotions that do not apply to rats (e.g. awe) and there are rat emotions that do not apply to human (e.g. fear of winged objects flying overhead). Similarly, AEs need not be human emotions at all. **They do however need to perform the same or similar *functions***. There may very well be AEs that humans and animals cannot relate to or even comprehend. Research is needed to identify, quantify and mechanistically characterize the functions of emotions in biological agents, and the extent to which they may be relevant for the design of AE in AI.

- Emotions are Dynamical. As many other emergent phenomena, they occur in more or less temporally stable fashion, depending on the properties of the underlying (often non-emotion related) computations. Animals are always in an emotional state, whether we have a name for it or not. Some biological emotions are readily identifiable across species or across individuals (e.g. fear) others may not be (e.g. happiness, awe). Emotions are not static, but follow trajectories, as the specific emotional responses to particular triggering stimuli arise and decay during the course of an emotional episode. Emotional expressions are possible endpoints of these trajectories, not the emotions themselves. Artificial Emotions should have the same characteristics. **Emotions transcend specific cognitive processes** (i.e. the fear of making a bad decision may *be mechanistically* the same as the fear elicited by the perception of a snake, although involving different brain circuits) and accompany them, as they unfold. Emotion processes are intertwined with non-emotional/cognitive processes.

- AI Lacks Memory 'Intrinsic Values', Artificial Emotions Can Help: In AI systems, all items memorized in a training set are usually valued the same way. This is well known not to be true in biology. Memory items have utility values (e.g. how often they are used or whether they are associated with positive or negative outcomes) computed dynamically, and/or intrinsic values (e.g. by their own nature, e.g. food). This value system is a constitutive part of the memory system, not an add-on, providing 'context' to a memory. This value system makes general memory storage, use and retrieval more efficient and more stable. It may be one of the reasons why animals do not need enormous amount of training data to learn and can learn emotionally salient (valuable) stimuli in a single trial. AI systems could significantly benefit from an intrinsic, **AE-mediated implementation of memory formation, storage and recall**.

- AI Should be Forgetful, AE Can Help*:* AI systems do not actively/intentionally forget. Animals do, as an important feature of their ability to function adaptively and efficiently. Forgetting is a necessary and useful function that is largely ignored in AI. **Emotions are one way (not the only way) to manipulate, sort, and reshape our memory landscape**. Artificial Emotions could/should do the same. Systems with such a feature will undoubtedly be of relevance to mental health (e.g. trauma, PTSD).

- Improving LLMs Using AE-driven Sentiment Analyses: LLMs use vast amount of data from the text corpora. This text data undoubtedly contains implicit and explicit emotional and value features. These are completely ignored and 'averaged'/'smoothed' over. LLM can be significantly improved by processing/using rather than eliminating or averaging the emotional value or content (e.g. functional sentiment analyses). Artificial Emotions provide a framework in which such analyses can be conducted, without relying on the intuitive (human) and often arbitrary understanding of what emotion a particular word carries.

- Improving AI Adaptive Capabilities Using AEs. Animals have specific bodies and brains shaped by evolution and their development is strongly shaped by the social milieu in which they mature. This milieu may itself be the product of cultural evolution and niche construction. An LLM, on the contrary, has a uniform and "bland" architecture whose parameters are almost entirely shaped by the training set. Unfortunately, the training set may have examples of behavior that range from the benevolent to the sadistic and given its size, cannot be manually annotated and filtered to eliminate inappropriate content. However, an AE system intrinsic to the LLM may provide the mechanisms required to **continuously adjust its training set** and thereby support the development of an individualized architecture, able to implement life-long learning and achieve appropriate adaptation to its physical and social environment and ethical behavior.

- Maladaptive Artificial Emotions: The price to pay for an efficient, robust, individualized emotion-coupled Artificial Intelligence is the possibility of emotional dysfunction. In biology, most mental disorders are accompanied by emotional dysfunctions and, conversely, emotional dysfunctions are almost always accompanied by cognitive or social impairments. While Artificial Emotions may introduce the risks of dysfunctions of the Artificial Intelligence system, **such 'maladaptive' AEs would represent a novel and valuable tool** to better understand maladaptive Biological Emotions and would support research aiming to identify and mitigate them.

- Proactive AI will Need AEs for Active Interactions and Collaborations: AI systems are increasingly and naturally being used as companions (e.g. LLMs, humanoid robots) or collaborators, whether they were designed for that purpose or not. Their ability to surprise the user and the ease of access to their vast knowledge base makes them useful tools for both social interactions and collaboration. Assistive AI is, and will continue to be, an important part of human technological future. It will also likely contribute to better mental health for some (e.g. elderly) and better therapy outcomes for others. Such systems need to recognize and express emotions, but will also need to understand them **using empathy and AE-rich internal models of others**. Furthermore, future AI systems (especially mobile robots) will involve robot-robot interactions and collaborations in complex and unpredictable environments (e.g.

autonomous vehicles on the road, not only robots on an assembly line). Such interactions will benefits from AEs to the same or even greater extent, that Biological Emotions have benefited human-human social interactions and collaborations. Importantly, currently most AI system are *passive and reactive*: they only engage after a stimulus or prompt. Proactive AI systems that *initiate* interactions and collaborations (e.g. LLMs that will spontaneously ask questions to improve their knowledge base irrespective of user interactions, robots that will ask humans for help in performing a task) are crucially lacking. New AI research is needed that will **allow AI to be proactive** in a safe and ethical manner and AEs are likely to play a critical role in this research.

- AEs for Predictive and Explainable AI. AI can generate complex predictions in often poorly understood manner. Explainable AI emerged as a way to address this issue but is still at its infancy. Most successful AI systems are beyond human-level explainable reach. Biological Emotions are to some extent responsible for one's future behaviors and recognizing emotions in others helps constrain our predictions of their future behaviors or lack thereof, and helps us plan for long term interactions with them. We cannot always explain how such predictions are generated (e.g. unconscious biases) but many tools exist in the clinical and psychological domains to deconstruct, analyze and neutralize these biases when they are identified as undesirable. Artificial Emotions could provide similar explanatory and predictive power, among other functions, which would **make autonomous AI more trustworthy, explainable, efficient and robust across multiple time scales**.

Acknowledgment: The authors wish to acknowledge the comments and edits of Dr. Michael Arbib.

Relevant References

Arbib MA, Fellous JM (2004) Emotions: from brain to robot. Trends Cogn Sci 8:554-561.

Barrett LF (2017) The theory of constructed emotion: an active inference account of interoception and categorization (vol 12, pg 17, 2017). Soc Cogn Affect Neur 12:1833-1833.

Fellous JM, Arbib MA (2005) Who Needs Emotions?: The Brain Meets the Robot. New York: Oxford University Press.

Fellous JM, Sapiro G, Rossi A, Mayberg H, Ferrante M (2019) Explainable Artificial Intelligence for Neuroscience: Behavioral Neurostimulation. Front Neurosci 13:1346.

Hudlicka E (2014) Affective BICA: Challenges and open questions. Biol Inspir Cogn Arc 7:98-125.

Hudlicka E (2020) The Case for Cognitive-Affective Architectures as Affective User Models in Behavioral Health Technologies. Lect Notes Artif Int 12197:191-206.

Hudlicka E (2023) Computational Models of Emotion and Cognition-Emotion Interaction. In: The Cambridge Handbook of Computational Cognitive Sciences (Sun R, ed), pp 973-1036. Cambridge: Cambridge University Press.

Koelsch S, Jacobs AM, Menninghaus W, Liebal K, Klann-Delius G, von Scheve C, Gebauer G (2015) The quartet theory of human emotions: An integrative and neurofunctional model. Phys Life Rev 13:1-27.

Ortony A, Clore GL, Collins A (1988) The Cognitive Structure of Emotions. Cambridge: Cambridge University Press.

Picard RW (2002) What does it mean for a computer to "have" emotions? In: Emotions in Humans and Artifacts (Trappl R, P. PP, Payr S, eds), pp 213-236. Cambridge, MA: MIT Press.

Picard RW (2003) Affective computing: Challenges. International Journal of Human-Computer Studies 59:55-64.

Scherer K (2010) The component process model: a blueprint for a comprehensive computational model of emotion. In: Blueprint for an affectively competent agents (Scherer KR, Banziger T, Roesch E, eds). Oxford, UK: Oxford University Press.

Scherer KR (2013) Emotion in Action, Interaction, Music, and Speech. Strungmann Forum Rep:107-139.

Sloman A, Chrisley R, Scheutz M (2005) The Architectural Basis of Affective States and Processes. In: Who Needs Emotions? (Fellous JM, Arbib MA, eds). Oxford, UK: Oxford University Press.