

Representing and Processing Emotions in a Cognitive Architecture

Pat Langley

Center for Design Research, Stanford University, Stanford, CA 94305 USA

Abstract

This paper briefly proposes a theory of emotions that clarifies their role in architectures for intelligent agents. The account posits that emotions take the form of symbolic cognitive structures, that generic emotional rules produce concrete instances of these relational concepts, that such rules underlie both generation and understanding of emotions, and that their results play a metacognitive role in influencing behavior. In this framework, emotions are central to high-level, human-like cognition.

Keywords

Emotional representation, Emotional processing, Cognitive architectures

Introduction

The topic of emotions has received little attention from the AI and cognitive science communities, at least compared to other phenomena. Emotions play a central role in most aspects of human life; they color and modulate our activities, both physical and mental. How are emotions related to cognition, and what function do they serve in a cognitive architecture? Science fiction often depicts human-level AI systems as devoid of emotion, but does this really make sense?

The traditional view is that emotions are ‘irrational’ holdovers from our evolutionary precursors. This perspective influenced much early AI work, which saw emotions as being detrimental to intelligent systems. We can build programs that – to some extent – reason, plan, and communicate without emotional components, but Simon (1967) argues that affect and emotion help control cognitive attention. And Damasio (1994) reports brain-damaged humans with little or no emotion who have difficulty making decisions. This suggests that human-level cognitive systems may actually *require* emotions.

Both academic papers and everyday language often confuse key concepts in this arena that are quite distinct. Here we propose four terms to denote different theoretical ideas:

- **Affect.** The valence and intensity for some experience.
- **Mood.** A global variant of affect for the entire cognitive system.
- **Emotion.** A relation among goals and beliefs for an event or object.
- **Feeling.** An affective or hormonal response associated with an emotion.

Third International Workshop on Human-Like Computing, September 28–30, 2022, Windsor Great Park, UK


✉ langley@stanford.edu (P. Langley)

🌐 <http://www.isle.org/~langley/> (P. Langley)

🆔 0000-0001-5260-7048 (P. Langley)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Lin, Spraragen, and Zyda (2012) introduce similar distinctions in their review of computational models within this area. A complete account would relate each factor to cognition, but here we focus on emotions, the most interesting from an architectural perspective. Many emotions are important enough to name: frustrated, regretful, disappointed, relieved, and even *schadenfreude*. Other mammals have emotions, but human variants are distinctive in their richness and complexity, which further suggests a strong cognitive component.

Elsewhere we discuss the need to incorporate emotions into cognitive architectures (Langley et al., 2009; Langley, 2017a). The account presented here builds on PUG (Langley et al. 2017), an architectural theory for embodied agents. This separates knowledge into concepts (which derive beliefs from perceptions), skills (which produce plans to achieve goals), and motives (which assign values to goals). Different aspects of emotions fit naturally into this framework.

One key question is whether emotions are better explained at the architectural level, and thus require specialized structures and procedures, or at the knowledge level, and thus require only the addition of content encoded in established representations. Another concerns what functions emotions can serve in intelligent agents, and thus what capabilities they support. In the sections that follow, we propose some tentative answers to these questions.

Representing Emotions

Before we can discuss processes that produce emotions, we must first consider how to represent them. Marsella et al. (2010) distinguish three frameworks: dimensional models (points in a continuous space); anatomical models (activations in neural circuits); and appraisal models (relations among cognitive structures). We focus here on the final framework, which lends itself to incorporation into theories of the human cognitive architecture (Langley et al., 2009).

Most dimensional models characterize emotions as points in a three-dimensional space: pleasure (measure of valence); arousal (level of affective activation); and dominance (measure of control). Synthetic characters often use such “PAD” models (e.g., Wachsmuth, 2008), but they ignore the fact that emotions are about some event, person, or object, and we can have mixed emotions about the same target. This suggests they involve much richer cognitive structures.

Appraisal models view emotions as inferred relations among mental structures based on situations. Ortony, Clore, and Collins (1988) describe 22 configurations that characterize emotions organized around events, objects, and other agents. These serve as ‘elicitation’ patterns on emotions that specify relations among an agent’s goals, intentions, expectations, and beliefs, as well as inferences about others’ mental states. Such emotional structures are *abstract* and *domain independent*, much like rules for dialogue (e.g., Gabaldon et al., 2014).

Our framework maps such “appraisal frames” onto *concepts* that reside in the PUG architectures’ long-term memory (Langley et al., 2016). Each conceptual rule defines some predicate that relates its arguments, much as in the Prolog formalism. The theory also maps instances of these generic structures onto *beliefs*, such as *resents(John, passed(Sam, CompSci101))*, that appear in working memory. Thus, we can encode emotional concepts and instances at the knowledge level, although they are atypical in that they can take other relations as arguments.

As an example, consider the familiar emotional concept *disappointed*, which relates a person *P* to some situation or event *S*. An instance of this emotion can arise when *P* wants *S* to become

true, P expected S to take place, and yet P believes that it did not actually occur:

- $disappointed(P, R) :- wants(P, R), expected(P, R), believes(P \text{ not}(R))$.

Similarly, we say that a person P is *jealous* of another person Q if P wants to possess an object S , believes that he does not possess S , and believes that Q does possess it. Complex emotions are specializations of basic ones that involve more conditions.

Although emotions have a clear structural element, their instances have an associated affective score that includes a valence and intensity, with the latter changing over time. Similarly, emotional concepts incorporate a value function that specifies this affect as a function of matched elements. Most appraisal theories include these features, but they do not map onto aspects of a cognitive architecture like PUG that includes them for other reasons.

Emotional Processing

Our theory also takes positions about how emotional structures are processed, which we can separate into two high-level cognitive tasks. One is *generation*, which produces emotions for the primary agent, such as $disappointed(John, failed(John, CompSci101))$. The other is *understanding*, which infers the emotions of other agents, such as $belief(John, disappointed(Jane, failed(John, CompSci101)))$. This maps onto the classic distinction between plan generation and plan understanding. Both appear necessary for a full account of emotion's relation to cognition.

We maintain that a single PUG mechanism, *conceptual inference*, underlies both generation of the primary agent's emotions and inference about those of others. This uses a variety of relational pattern matching like that in logic programming, although an alternative would use abduction rather than deduction. This draws on processes available in the existing architecture, which offers further evidence that emotions are knowledge-level phenomena.

However, this raises a question about why affective scores are more intense for our own emotions than those inferred for others. Presumably, this difference results from coefficients in the value function associated with each emotional concept, which will be higher for the primary agent and lower for inferences about other agents. This may also explain why memories of past emotions are typically less intense than the original experience. This suggests that the distinction between 'hot' and 'cold' emotions is not a dichotomy but rather a continuum.

A full theory should clarify not only how the architecture generates and infers emotions, but also how they influence other cognitive processing. Thus, it should specify how emotional instances impact the agent's physical behavior (e.g., crying about loss or punching someone) or its cognitive processing (e.g., changing goal priorities or invoking planning). This supports the view that emotions are not evolutionary relics, but rather high-level regulators of cognition.

We postulate that this relates to people's ability to think about thinking – *metacognition*. Recall that emotional concepts specify abstract relations among goals, beliefs, and expectations, which are matched during conceptual inference. This suggests a promising hypothesis:

- *Emotions play a **metacognitive** role that operates over and influences base-level cognition.*

That is, emotional processing inspects traces of basic cognition and alters its course in response. This view follows Simon (1967) in claiming that emotions play a regulatory role in cognitive attention, but it also suggests the need for additional mediating structures.

The PUG architecture already incorporates long-term structures called *motives* that match against domain-level beliefs, generate top-level goals, and compute the latter's values. Elsewhere, we have proposed motives that match against emotions and generate goals which may involve changing emotions of the primary agent or others (Langley, 2017b). For instance, the rule

- $disappointed(P, R), believes(P, cause(Q, R)) \rightarrow wants(P, disappointed(Q, -))$.

encodes an 'eye for an eye' motive, so that if someone believes another agent caused something that disappointed him, then he desires to reciprocate in kind. The priority of this goal will depend on the motive's value function. We also conjectured that such motives make up the agent's personality, reflecting stable, domain-independent regularities in cognition and behavior.

Discussion

Our theory of cognition and emotion draws on ideas that have appeared elsewhere in the literature, many of them discussed by Marsella, Gratch, and Petta (2010) and by Lin, Spraragen, and Zyda (2012). Both Sloman (2001) and Minsky (2007) emphasize that emotion and cognition are closely intertwined, but they offer few details. Gratch and Marsella (2004), Marinier, Laird, and Lewis (2009), and Hudlicka (2007) describe detailed appraisal models of emotions embedded in cognitive architectures, including their modulation of cognition and their intensities.

One notable difference is that their accounts assume a fixed set of appraisal frames, each associated with a distinct emotion, whereas our theory allows an arbitrary number of emotional concepts, some defined in terms of others. In this sense, it comes closer to Gordon and Hobbs' (2017) analysis, which also defines emotions in relational logic. Our approach benefits from PUG's theoretical distinction between concepts and skills, as does its reliance on the architecture's motives to explain emotions' metacognitive effects.

In summary, our computational account of emotion brings together ideas from appraisal theory and cognitive architectures. The most important claims of the framework, which it shares with some other work that follows similar lines, are that:

- Emotions are symbolic cognitive structures with numeric annotations;
- Generic emotional rules generate specific instances of these concepts;
- These rules are used to generate emotions and to infer those of others;
- Emotions play a metacognitive role in influencing cognition and behavior;
- These influences are mediated by generic motives that assign values to goals.

In this theory, emotions are not irrational vestiges of evolution. Rather, they are linked directly to structures and processes in the cognitive architecture, specifically ones in the PUG framework, which should require only minor extensions to incorporate them.

We maintain that PUG's division of knowledge into concepts, skills, and motives makes it especially suitable for explaining emotions and their relation to both cognition and personality. However, we must still demonstrate the architecture's ability to reproduce emotion-related phenomena in realistic scenarios that are similar to ones humans encounter. We must also work out details about how to calculate emotional intensities as a function of time, different agents, and other factors, but we remain optimistic about the framework's potential in this arena.

Acknowledgments

The research reported here was supported by Grant No. FA9550-20-1-0130 from the US Air Force Office of Scientific Research, which is not responsible for its contents. We thank John Laird and the reviewers for constructive comments that improved the paper.

References

- Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Putnam.
- Gabalton, A., Langley, P., & Meadows, B. (2014). Integrating meta-level and domain-level knowledge for task-oriented dialogue. *Advances in Cognitive Systems*, 3, 201–219.
- Gordon, A. S., & Hobbs, J. R. (2017). *A formal theory of commonsense psychology: How people think people think*, 513–544. New York: Cambridge University Press.
- Gratch, J., & Marsella, S. (2004). A domain-independent framework for modeling emotion. *Cognitive Systems Research*, 5, 269–306.
- Hudlicka, E. (2007). Reasons for emotions: Modeling emotions in integrated cognitive systems. In W. Gray (Ed.), *Integrated models of cognitive systems*. New York: Oxford University Press.
- Langley, P. (2017a). Progress and challenges in research on cognitive architectures. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (pp. 4870–4876). San Francisco: AAAI Press.
- Langley, P. (2017b). A cognitive systems analysis of personality and conversational style. *Proceedings of the Fifth Annual Conference on Cognitive Systems*. Troy, NY.
- Langley, P., Barley, M., Meadows, B., Choi, D., & Katz, E. P. (2016). Goals, utilities, and mental simulation in continuous planning. *Proceedings of the Fourth Annual Conference on Cognitive Systems*. Evanston, IL.
- Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, 10, 141–160.
- Lin, J., Spraragen, M., & Zyda, M. (2012). Computational models of emotion and cognition. *Advances in Cognitive Systems*, 2, 59–76.
- Marinier, R., Laird, J., & Lewis, R. (2009). A computational unification of cognitive behavior and emotion. *Cognitive Systems Research*, 10, 48–69.
- Marsella, S., Gratch, J., & Petta, P. (2010). Computational models of emotion. In K. R. Scherer, T. Banziger, & E. B. Roesch (Eds.), *A blueprint for affective computing: A sourcebook and manual*. Oxford: Oxford University Press.
- Minsky, M. (2007). *The Emotion Machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. New York: Simon and Schuster.
- Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. New York: Cambridge University Press.
- Simon, H. A. (1967). Motivational and emotional controls of cognition. *Psychological Review*, 74, 29–39.
- Sloman, A. (2001). Varieties of affect and the CogAff architecture schema. *Proceedings of the AISB'01 Symposium on Emotion, Cognition, and Affective Computing* (pp. 39–48). York, UK.
- Wachsmuth, I., Lenzen, M., & Knoblich, G. (2008). *Embodied communication in humans and machines*. Oxford, UK: Oxford University Press.