Put Feeling into Cognitive Models: A Computational Theory of Feeling

Robert L. West (robert.west@carleton.ca),
Brendan Conway-Smith (brendan.conwaysmith@carleton.ca)
Institute of Cognitive Science, Carleton University, Ottawa, ON K1S5B6 Canada

Abstract

Feelings are potentially conscious experiences that inform us about brain/body states related to drives (e.g., feeling hungry), emotions (e.g., feeling angry) and knowledge states (e.g., feeling unsure). In this paper we propose a unified computational definition of feelings that can be used to add feelings to cognitive models.

Accounting for feelings in cognitive models is important since feelings have strong effects on human performance and decision-making. However, there is considerable disagreement over what feelings are and how, or if, they can be incorporated into cognitive models. We address this issue by providing a functional, computational definition of feelings.

Computational Theory of Mind (CTM) is an area of philosophy that argues that the brain is a form of computer. There are a variety of arguments in favour of this view, likely the most well known belong to Fodor (2000). Likewise, there are a variety of criticisms of this view, probably the most well known are Searle's (1984). Internally, CMT theorists argue about the right way to map computation to cognition. Mostly these discussions revolve around knowledge and language, but the question of how to relate feelings to computation has been broached, so this is one source of ideas about how to computationally implement feelings.

Another source of ideas is Cognitive Modeling itself. Cognitive modeling can be considered an empirical endeavour. The goal of Cognitive Modeling is to use computational modeling to represent cognitive theories and to test these theories through comparisons to data. The ultimate goal of Cognitive Modeling is to build a Unified Cognitive Architecture capable of simulating all or most human cognitive abilities (Newell, 1990). Cognitive architectures, such as ACT-R (Anderson & Lebiere, 1998) and SOAR (Laird, 2012) have been very successful in modeling knowledge driven behaviour but it is not clear how to add feelings to these architectures. However, by examining the structure of these architectures, the options for adding feelings can be elucidated.

What are Feelings?

CTM debates are focused around concepts such as symbolic representation, referents, semantics, propositions, qualia, and meaning. CTM is intended to describe the relationship between computation and the brain, but because most of the discussion is based around knowledge and language, it is unclear if these concepts are meant to apply beyond this domain (Rescorla, 2015). In particular, there seems to be an intuition that feelings are not the same as thought or language, and so must be computationally represented in a different way.

According to Damasio (2019) feelings are mental representations of non-symbolic bodily states, which are used for decision making. According to Alston (1969) feelings are, "spontaneously-emerging occurrent phenomenal experiences," which he refers to as "datable states of consciousness." However, Arango-Muñoz & Michaelian (2014) indicate that feelings do not involve "properly propositional content."

Overall, there seems to be agreement that a feeling is a unitary phenomena that we have potential conscious awareness of. Feelings can factor into decision making but there is an intuition that feelings are somehow different from propositional, symbolic knowledge. Finally, feelings are derived from more complex, distributed phenomena, such as emotions and drives.

Noetic Feelings

In addition to drives and emotions, feelings can also be derived from states of knowing or learning. These feelings have been referred to by terms such as, feelings of knowing or FOK (Hart, 1965), metamemory (Flavell, 1971), knowledge judgements (Schneider, 2000), cognitive emotions (Standish, 1992), and epistemic feelings (Arango-Muñoz & Michaelian, 2014). In the following paper we will subsume this lexicon under the term "noetic feelings." This follows Metcalfe's (2013) identification of "noetic" to mean cognitive phenomena in which the referent concerns an internal state or internal representation.

Research indicates that noetic feelings drive memory search as subjects take more time to search their memory if they "feel" they know it (Barnes et al., 1999). Studies also show that noetic feelings are a reliable signal of the likelihood of memory retrieval (Hart, 1965), and feelings of probable retrieval success or retrieval failure affect the strategy used to engage the problem (Conway, 2009; Singer & Tiede, 2008). Noetic

feelings have also been reliably correlated with improved learning outcomes (Wang, Haertel, & Walberg, 1990). Subjects will also spend more time learning words previously considered to be difficult to remember (Nelson & Leonesio, 1988). Moreover, the "feeling of rightness" has been studied in the rapid solving of complex, real-world problems (Thompson et al., 2011).

Models associated with noetic phenomena include Reder's (1996) use of the source of activation confusion (SAC) model, Dougherty's (2001) multiple-trace memory model, Metcalfe's (1993) holographic associative model, and Sikström and Jönsson's (2005) stochastic dri model of memory strength to explain delayed judgement of learning.

Thus research indicates that noetic feelings are a guidance system integral to directing cognitive processes. Progress toward accurately describing human cognition requires integrating noetic feelings into cognitive modeling.

Reasoning from Architectures

There is a tendency in CTM papers to focus on foundational issues. In the case of emotions, for example, this manifests as a concern over establishing what emotions are before considering how they can be computationally represented. For example, emotions are defined variously as bodily states (Damasio, 2019), perception (Prinz, 2006), and natural kinds (Barrett, 2006). However, since there is no agreement on the status of emotions we have no foundational basis to reason about the nature of the feelings that are derived from emotions.

In contrast to this foundational approach, we ground our work on the function of feelings within cognitive models. That is, we take a top down functional approach as opposed to a bottom up foundational approach. Ideally, these two different approaches can inform each other, but it is important not to confound the two.

We take the ability of cognitive models to account for data as evidence that the model embodies something true about the computational functionality of the brain. One criticism that can be levelled at this approach is that there are many different cognitive models. However, our focus is not on the differences between the models, but rather on their similarities. We argue that there are significant convergences in cognitive modeling at the level of the architecture. More generally, we interpret unified cognitive architectures as a way of grounding theory in functional coherence, without engaging with foundational issues. In particular, we focus on the Common Model Architecture.

Common Model of Cognition

The Common Model of Cognition (formerly known as the Standard Model of Cognition) is a conceptual architecture put forward by Laird et al. (2017). The concept of a common model is based on Laird et al.'s claim that there has been significant convergence across cognitive architectures over time, to where we are now at the point that we can talk about a common cognitive architecture. The common model describes a conceptual architecture that is common to most, if not all, cognitive architectures capable of modelling complex human behaviour.

The basic structure of the common model is shown in Figure 1. The common model describes a production system (corresponding to procedural memory) that interacts with different modules through a buffer system that corresponds to working memory. The architecture is parallel and asynchronous, with the production system acting as a control system. There are significant divergences in terms of how components are implemented in different common model-type architectures (e.g., spiking neurons, neural networks, high dimensional vectors, semantic networks, Bayesian networks, graph theory, etc.). However, the common model describes the common functionality across different implementations.

The common model is not meant to describe all of human cognition, it is a model of cognitive control and decision making. As Newell (1990) noted, this is one possible starting point for understanding cognition. In contrast, CTM appears to have knowledge and language as its starting point. In other words, CTM is based on *knowing* while the common model is based on *doing*. Bridging the two is conceptually tricky, not least because they use the same terms in different ways. In this paper we will attempt to merge CTM work on feelings with the common model. Specifically, we argue that feelings are best modelled as non propositional representations in buffers (related to this see West & Young, 2017, for a discussion of representing amygdala states in the buffers).

Qualia

Qualia refers to the qualitative differences between our conscious experience of thoughts, senses, emotions, and drives. Explaining how different patterns of neural activity can produce these qualitatively different experiences is part of what Chalmers (1996) referred to as the hard problem of consciousness (Chalmers, 1996). There continues to be much debate on the subject and

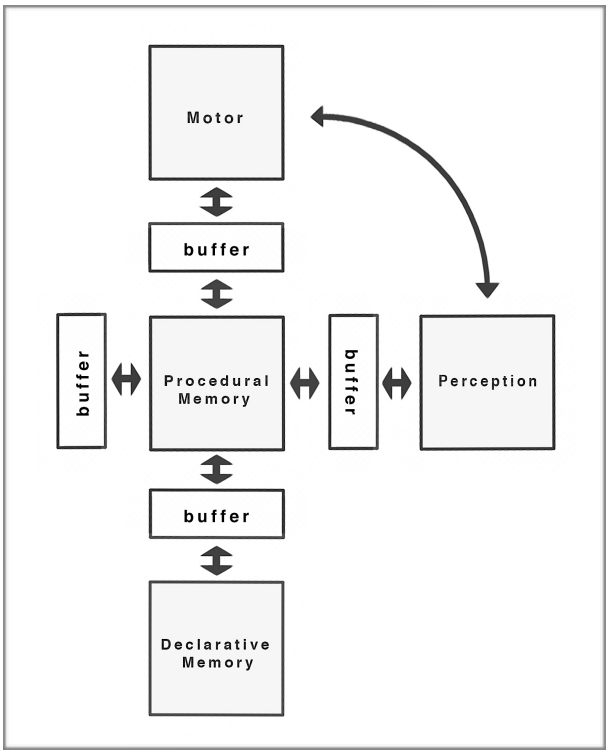

Figure 1. The Common Model of Cognition

the matter is far from settled. Given that there is no agreed upon definition of consciousness we will not make any strong claims about what parts of the common model are conscious. However, because the architecture can report the buffer contents it is clear that the buffer contents are potentially conscious, and are definitely conscious when reported. For example, if a model reports remembering a brown cat this would correspond to a conscious awareness of this cat in memory. Likewise, feelings can be consciously experienced but are not necessarily always consciously experienced (Redder & Schunn, 1996; Metcalfe & Son, 2012; Son & Kornell, 2005).

Modules and mechanisms

Fodor (1983) has greatly influenced how to think about the brain in terms of modules. However, cognitive modellers almost universally ignore Fodor's foundational requirements for modularity. So we will not follow Fodor on this. Instead we take modules in cognitive models to represent mechanisms in the brain (Betchel, 1994).

The mapping between a module in a model and a module in the brain can occur in different ways. It could be one to one, where a module in the architecture maps directly to an area of the brain, as suggested by J. Anderson (e.g.; Anderson et al., 2004). In terms of emotion, this approach is represented by theories of basic emotions, which postulate distinct neural modules for processing specific emotions. For example, (Panksepp, 1998) postulates areas for basic emotions based on comparisons of mammalian brains.

However, a computational module could also map to a network of multipurpose modules assembled to generate a higher level function, as suggested by M. Anderson (2010). In terms of emotion, this approach is represented by theories of complex emotions and emotional networks. Such networks could produce complex sets of feelings or blended feelings.

Brain wide states, such as neurotransmitter levels or hormones can also be modeled. For example, Ritter et al. (2006) describe a system for the ACT-R architecture that specifies the effects of hormones and neurotransmitter levels on modules in the architecture. Likewise, Core Affect theory (Russell et al, 1999) models brain wide chemical states in terms of a two dimensional valence/arousal model. We assume that brain wide states contribute to feelings through their impact on modules.

Propositions, Symbols and Feelings

Cognitive models either use symbolic propositional knowledge or, in the case of neural networks and spiking neuron models, they act as if they do. This makes sense for modelling knowledge driven processing but it raises a concern because consciously experienced feelings seem to possess qualia without associated propositional content. This can evoke complex philosophical questions (as represented by the thought experiment of Mary the colour blind scientist). However, we hope to avoid questions related to qualia by focusing on function. We begin by considering if the role of buffer representations is necessarily symbolic or propositional.

The buffer contents could be considered as representing propositional knowledge in the sense that the buffers are considered to contain true information. The buffer contents could also be considered to be symbolic representations in that they can correspond to things in the real world. However, what the buffers actually contain is the outputs of a module. How this relates to the state of the world is dependent on the relationship between the module and the world. If we consider the immediate function of the buffer contents for choice or decision making, they do not refer to anything except the matching code in the *if* condition of a production rule.

Whether or not the buffer contents should be considered propositional or symbolic is hard to answer because there is very little agreement on how to define

these terms. Many people (but not all) would agree that the linguistic representation of the statement "there is a black cat" is both propositional and symbolic. However, if we change it to a visual representation of a black cat then some would argue that it is no longer propositional or symbolic, while others would maintain that this changes nothing.

The key is to distinguish between the status of a representation conferred by being in a buffer versus the status of a representation conferred by its syntactic or representational structure. We argue that being in a buffer does not directly imply that a representation is symbolic or propositional as the only essential requirement is that the code in a buffer can match the code in the *if* part of a production. What could potentially distinguish a buffer containing a feeling from a buffer containing knowledge is the computational structure of the representation itself.

Here it is important to note that buffer contents in the brain are represented by neural firing patterns. However, these neural patterns can be represented by a symbol in a model without implying that the pattern has a symbolic function in the brain. For example, if the average spiking rate of a group of neurons was expressed as 42, although 42 is a symbol, it is without meaning unless you know the question that it answers. Even the numerical value of 42 is meaningless without knowing the units of measurement. Using symbols in a model may be merely a convenience for the modeller, it does not necessarily imply a theoretical commitment. For example, if the feeling of anger was represented by putting the word “anger” in a buffer, this would not imply anything.

Following this we can ask — what would it mean if a buffer contained a word (or neural pattern) corresponding to a qualia, such as anger, or tired, or unsure? Functionally, because feelings can be experienced consciously, we know from experience that we use them to make decisions. Whether or not they are propositional or symbolic seems to depend on the extent of the conceptual framing of the decision. If it is simply the moment of matching to a production then it can be argued that they are neither propositional nor symbolic. If the question is “why did you hit that man?” then the function of the feelings involved could be argued to be propositional and symbolic, in terms of their role in larger decision.

Mentalese

Mentalese is a concept invented by Fodor (2000) to distinguish between language and the language of thought. However, we use the term in the broader sense outlined by Pinker (1997), in which different modalities have their own mentalese. For example, Pinker proposed that we have visual mentalese. We interpret the contents of the buffers to be mentalese and propose there are different types of mentalese. The implication of this is that the mentalese used in one buffer may not be directly translatable to the mentalese used in another.

This is an important issue for the common model. If two buffers use the same mentalese, then a single production can transfer information directly from one buffer to another without reference to the content, but if they use different mentalese there needs to be a translation. Minimally, this would require a different production for each object of translation. For example, if a representation of a stop sign is in the visual buffer, to put a representation of “stop” in the goal buffer requires a production recognizing the visual mentalese representation of the stop sign on the *if* side and, on the *then* side, puts a goal mentalese representation of “stop” in the goal buffer. Alternatively, it is possible that there is a common mentalese for knowledge and the different modalities translate information into this common language before it arrives in the buffer. Most common model models are programmed as if the second option is true. Ideally, it will be possible to empirically answer this question.

However, our common experience with feelings indicates that, although we can label them, we often have difficulty putting them into words. The entire field of poetry is arguably dedicated to this effort. Another distinguishing factor is that we cannot alter our feelings in the same way we can alter our knowledge or goals. For example, if I have stopped at a stop sign and there is no traffic, I can quickly alter the content of my goal buffer from “stop” to “go.” In contrast, if I am angry and I realize that it is unwise to be angry, I cannot simply change the feeling in the buffer to another emotion. These examples suggest that feelings have their own mentalese and that the production system cannot directly alter this mentalese. Combined with the fact that some people have difficulty labeling their feelings, this suggests that the production system learns, through experience, to associate knowledge-mentalese labels with different feeling-mentalese representations. This suggests that, feeling-mentalese functions more like a sign system, similar to what animals are capable of.

Feelings are also associated with phenomena such as facial expressions and hormonal release. However, at the 50 millisecond time scale of productions we are talking about choice. For example, in an approach avoidance scenario such as a monkey contemplating food left out in a clearing, we simultaneously experience the feelings of hunger for food and fear of predators in the clearing. The result is a vacillating,

back and forth behavior accounted for by opposing productions firing back and forth.

Related to this, mindfulness training in Cognitive Therapy can be understood as learning to translate feelings to knowledge in order to use the more advanced properties of knowledge to gain a better purchase on our behavior. Once a feeling is labeled it has been translated to knowledge, but this new representation is not a feeling, and our experience tells us that the feeling still independently exists. For example, if you are walking home in the dark after a watching a vampire movie, you might experience fear. By translating the fear-feeling to knowledge, you can reason that vampires are not real and so you are not in danger. However, while this will help, and may decrease the fear feeling, the fear feeling will independently persist in the short term.

Feelings as Metadata

We propose that feelings are metadata and that feeling-mentalese is a language appropriate for expressing metadata, whereas knowledge-mentalese is a language appropriate for expressing knowledge. This makes sense since we know that, computationally, metadata expressions are typically different from knowledge expressions. For example, metadata is often best expressed through statistics and high dimensional spaces, whereas knowledge is often best expressed through propositional statements and logical operators. This also accounts for the fuzzy, non-verbal qualia of feelings.

To maintain the distinction between knowledge and metadata, we argue that knowledge statements about feelings, such as, I feel angry, or, I feel confused, are translations performed by productions that recognize metadata states and create knowledge based statements about them. So, as such, these statements are knowledge and not feelings. Questions about whether feeling-metadata can be considered propositional or symbolic, we believe, depends on how the data is used in the model.

Another computational distinction we think we can make is that feelings are bottom-up, read-only statements. That is, feeling-representations are placed in buffers by their associated modules and the central production system cannot alter them. Only the module that created them can alter them. The production system may or may not have direct access to the module. In contrast, knowledge representations in the buffers can be altered directly by the production system, as is common in common model architectures.

Conclusion

We have presented a computational theory of feelings based on the common model architecture. More broadly, we have shown how cognitive architectures can be applied to clarify philosophical issues, particularly in CTM. We believe this type of work is important as conceptual confusion over issues, such as the difference between knowledge and feelings, can conceptually impede the creation or acceptance of cognitive models involving these phenomena. Finally, by stating our ideas in terms of a cognitive architecture we have made them computationally unambiguous. Other, different models are possible, but they should be stated clearly, in computational terms, and grounded in a viable cognitive architecture.

References

Arango-Muñoz, S., Michaelian, K. (2014). Epistemic Feelings and Epistemic Emotions (Focus Section). Philosophical Inquiries.

Anderson, J. & Lebiere, C. (1998). The atomic components of thought. Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, J., Qin, Y., Stenger, V., & Carter, C. (2004). The relationship of three cortical regions to an information-processing model. Journal of Cognitive Neuroscience, 16 (4), 637-653.

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. Behavioral and Brain Sciences, 33(4), 245–266.

Alston, William P. Feelings. The Philosophical Review 78(1): 3-34, 1969.

Barnes, A., Nelson, T., Dunlosky, J., Mazzoni, G., & Narens, L. (1999). An integrative system of metamemory components involved in retrieval. Attention And Performance Xvii–Cognitive Regulation of Performance: Interaction of Theory And Application, 17, 287–313.

Barrett, L. F. (2006). Are Emotions Natural Kinds? Perspectives on Psychological Science, 1(1), 28–58.

Bechtel, W. (1994). Levels of Description and Explanation in Cognitive Science. Minds and Machines, 4, 1-25.

Chalmers, David J. (1996). The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.

Computational Theory of Mind. (2015). Stanford encyclopedia of philosophy. Retrieved from https://plato.stanford.edu

Conway M.A. (2009). Episodic memories. Neuro psychologia. 47: 2305-13. PMID

Damasio, A. (2015). Scientific American 2019. Oct. 16 Nation, 294(20), 11-18. Retrieved from MAS Ultra.

de Sousa, R. Epistemic feelings. In: Georg Brun, Ulvi Doˇguoˇglu, and Dominque Kuenzle, eds., Epistemology and emotions. Ashgate, 2008.

Dougherty, M. R. P. (2001). Integration of the ecological and error models of overcon dence using a multiple-trace memory model. Journal of Experimental Psychology: General, 130, 579–599.

Flavell, J.H. (1971). What is memory development the development of? Human Development, 14, 272-278.

Fodor, J. (1983). Modularity of Mind: An Essay on Faculty Psychology. Cambridge, Massachusetts: MIT Press.

Fodor, J. (2000). The Mind Doesn't Work That Way:The Scope and Limits of Computational Psychology.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. Journal of Educational Psychology.

Laird, J. & Lebiere, C. & Rosenbloom, P. (2017). A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. AI Magazine. 38.

Laird, J. E. (2012). The SOAR cognitive architecture. MIT Press.

Metcalfe, J. (1993). Novelty monitoring, metacognition, and control in a composite holographic associative recall model: Implications for Korsako amnesia. Psychological Review, 100, 3–22.

Metcalfe, J., & Son, L. (2013). Anoetic, noetic, and autonoetic metacognition. Foundations of Metacognition.

Metcalfe, J., & Son, L. K. (2012). Anoetic, noetic, and autonoetic metacognition. In M. Beran, J. L. Brandl, J. Perner, and J. Proust (Eds.). Foundations of Metacognition (pp. 289-301). Oxford University Press

Metcalfe, Janet & Son, Lisa. (2013). Anoetic, noetic, and autonoetic metacognition. Foundations of Metacognition

Nelson, T. O., & Narens, L. (1990). Metamemory: a theoretical framework and new ndings. In G. H. Bower (Ed.), e psychology of learning and motivation (Vol. 26, pp. 125–173). New York: Academic Press.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe, & A. J. Shimamura (Eds.), Metacognition: Knowing about knowing (pp. 1–26). Cambridge, MA: MIT Press.

Newell, A. (1990). Unified theories of cognition. Cambridge: Harvard University Press.

Panksepp, J. (1998). Affective neuroscience: The foundations of human and animal emotions. New York: Oxford University Press.

Pinker, S. (1997). How the Mind Works. New York: Norton.

Prinz, J. (2006). Canadian Journal of Philosophy, Volume 36, Supplement Col. 32, pp. 137-160.

Proust, Joelle (2009). Is there a sense of agency for thought? In Lucy O'Brien & Matthew Soteriou (eds.), Mental Actions. Oxford University Press.

Reder, L. , & Ritter, F. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. journal of Experimental Psychology: Learning, Memory, and Cognition, 18, 435-451.

Reder, L. M., & Schunn, C. D. (1996). Metacognition does not imply awareness: Strategy choice is governed by implicit learning and memory. In L. M. Reder (Ed.), Implicit memory and metacognition (pp. 79–122).

Ritter, F. E., Reifers, A. L., Klein, A. C., & Schoelles, M. J. 2006. Lessons from Defining Theories of Stress. In W. Gray (Ed.) Inte- grated Models of Cognitive Systems. New York, NY: Oxford University Press.

Russell, James A.; Barrett, Lisa Feldman (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant" (PDF). Journal of Personality and Social Psychology. 76 (5): 805–819.

Searle, J. (1984), Minds, Brains and Science: The 1984 Reith Lectures, Harvard University Press.

Schneider, W., Visé, M., Lockl, K., Nelson, T. (2000). Developmental trends in children's memory monitoring - Evidence from a judgment-of-learning task. Cognitive Development. 15. 115-134.

Singer, M., & Tiede, H. L. (2008). Feeling of knowing and duration of unsuccessful memory search. Memory and Cognition, 36, 588-597.

Son, L. K., & Kornell, N. (2005). Meta-confidence judgments in rhesus macaques: Explicit versus implicit mechanisms. In Terrace, H.S. & Metcalfe, J. (Eds.), The Missing Link in Cognition: Origins of Self-Knowing Consciousness. Oxford University Press.

Standish, P. (1992), In Praise of the Cognitive Emotions. Journal of Philosophy of Education, 26: 117-119. opmental inquiry. American Psychologist, 34, 906–911.

Strawson, P. F. (1948). Truth. Analysis 9 (6): 83-97.

Sikström, S., & Jönsson, F. (2005). A model for stochastic dri in memory strength to account for judgments of learning. Psychological Review, 112, 932–950.

Thompson VA, Prowse-Turner J, Pennycook G. (2011) Intuition, reason & metacognition. Cogn Psychol 63:107–140

Wang, M., Haertel, G., & Walberg, H. (1990). What Influences Learning? A Content Analysis of Review Literature. The Journal of Ed. Research, 84(1), 30-43.

West, R. L., & Young, J. (2017). Proposal to add emotion to the standard model. 2017 AAAI Fall Symposium Technical Report Volume 17, Symposium 6: A Standard Model of the Mind