Functions, Creatures, Learning, Emotion

Stephen Petersen[*]
Department of Philosophy
University of Michigan
2215 Angell Hall
435 South State St.
Ann Arbor MI 48104
spetey@umich.edu

Abstract

I propose a conceptual framework for emotions according to which they are best understood as the feedback mechanism a creature possesses in virtue of its function to learn. More specifically, emotions can be neatly modeled as a measure of harmony in a certain kind of constraint satisfaction problem. This measure can be used as error for weight adjustment (learning) in an unsupervised connectionist network.

As might be expected of a philosopher, I don't have empirical results; instead, I'll briefly suggest here how an information-theoretic approach to emotions might fit into a broader conceptual framework. I can only hope these thoughts will help inform the fascinating science that is being done in this area, about which I am still only learning.

Functions and Creatures

Intuitively, many things in the world have *functions*—they are "supposed to" do something. Hammers are supposed to drive nails, and hearts are supposed to pump blood. Something has a function, roughly, if it was designed to produce a certain effect. This design process, in turn, is basically one of feedback, encouraging some effects and discouraging others. The design can be a result of intelligence, as when we build a better mousetrap. Alternatively design can arise from a non-intelligent process like natural selection.[1]

I use this popular philosophical notion of function to characterize my own idiosyncratic notion of a *creature*. To be a creature, roughly, is to possess at least one function that can be performed autonomously—that is, solely through the performance of sub-functions.[2] The intuition is that creatures are things with "wants" (broadly speaking) in the world

that they "try" to achieve through their sub-functions. Plants want their leaves aimed at the sun, robins want to stay warm in winter, and humans want better local theater.

My characterization has what may seem a strange consequence: a thermostat counts as a creature. A thermostat has a function to regulate room temperature, which it does autonomously through its internal function to correlate a wire coil and a heater switch. In virtue of this internal function it "wants", loosely speaking, for the room to be some temperature. I embrace this as the natural result of two advantages of my characterization: first, it allows creatures to be mechanical or biological—it is not chauvinist about who can be intelligent. Second, it captures our intuition that intelligence is a matter of degree on a very wide spectrum. Naturally the thermostat is about as simple and unintelligent as a creature can be, but it is a creature by my account nonetheless.

Creatures and Learning

A particularly sophisticated creature, unlike our thermostat, will have sub-functions designed to alter other sub-functions in the direction of fulfilling its more basic functions. A mouse, for example, can alter its cognitive dispositions to behavior when it runs a maze a few times. A function to adjust cognitive functions is just a capacity to learn. Learning allows creatures to adapt better to local environments for the better achievement of their goals—in other words, on a popular account of intelligence, learning allows for greater intelligence.

Put another way, a learning function designs some other cognitive function by giving it feedback, inhibiting bad effects and encouraging good effects. That means, in turn, that the learning function serves as an internal measure of error for the system. This measure of error implicitly contains both a representation of the goal state the creature would "like" to be in, and a representation of the creature's current state with respect to that goal. That is to say, creatures that learn have both *conations* (desire-like states) and *cognitions* (belief-like states). The general form of unsupervised learning involves, then, a matching of conations to cognitions. A creature is functioning better as a self-teacher, you might say, to the extent that it can minimize these internal error measures. The process of learning can be described as a creature's attempting to match its cognitions to its conations—attempting to match its representation of how

---

[1]This is a loose but neutral characterization of some more precise philosophical approaches to functions, which turn out to be surprisingly tricky to characterize. See (Manning 1998) for a brief, helpful overview.

[2]I emphasize that this is rough, especially the 'autonomy' clause. But I think this is okay; I only lean on the functional approach.

things are to its representation of how it would "like" things to be.

Let me briefly indicate how this could be implemented computationally. The creature will have, in virtue of its design, various hardwired conative and cognitive mechanisms. A mouse cannot easily learn away its desire for food, nor can it easily learn away its basic visual inputs. A complex creature with many sub-functions may at any time have conflicting hardwired or learned demands that need resolving. Each error measure—each pair of cognition and conation—becomes a soft constraint, no one of which is decisive, but each one of which has some claim on revising the creature's thinking system. Represent these cognition and conation pairs as nodes in a connectionist network with positive weight between them. Furthermore some of the cognitions, and some other of the conations, will have default activation from the creature's basic hardwiring endowed it by the world—this can be represented in a network with a positive connection to a strongly activated "world node". In the mouse, for example, a desire for food and a retinal cell impulse would have such strong default activation. Of course this default activation can be overruled; any one conation or cognition can be given up for a balance of satisfying the others. The network then calculates behavior, given inputs, as a coherence problem.[3] Learning, finally, can be modeled as a weight-adjustment problem (within externally determined constraints) toward maximizing the overall potential coherence of the system.

Learning and Emotion

The account of emotions that falls out is simple: emotions are something like reports of the coherence levels between cognitions and conations, for the purpose of feedback and learning.

Let me try to make this more plausible. An intelligent agent will have, at any time, a degree of positive or negative desire for p (represented as activation on an interval of $[-1, 1]$) and a correlated degree of positive or negative belief that p. When many of its desires match many of its beliefs in level of activation, it is intuitively "happy". Things are as the creature "wants" them to be, and so there should be positive reinforcement of whatever has led to this state. Notice that creatures can be happy (in the sense of positive affect) even when in *fact* they are being deceived, and their desires are not satisfied. They are happy because their desires *seem* to be satisfied; their cognitive representations match their conative ones. On the other hand, a thinker with many mismatches between beliefs that p and desires that p will thereby feel compelled to change its situation. There will be internal, negative reinforcement for what led to the current state. It seems natural to say such a creature is sad, or anxious, or both.

The result is much like the *cognitive appraisal theory* of

emotions, according to which emotions are (or are the result of) a comparison of cognitive and conative appraisals of a situation.[4] But according to the proposed version, these "cognitive appraisals" do not have to occur at the *propositional* level. Emotions can result from the cognitive proprioceptions of low-level bodily processes as well, as work such as Antonio Damasio's would have it.[5] And the proposal can also explain in part why emotions are so central to decision-making. According to the proposal, emotions amount to feedback about how the agent is doing. Without such feedback, it's understandable why an agent like Phineas Gage had a hard time attaining his desires.

Though computational aspects build on work by Thagard, the theoretical foundations of my account make for different implications, especially when it comes to emotions. Thagard's model of "emotional coherence" in HOTCO is different in kind from his other models.[6] HOTCO requires adding separate "valence" values for nodes in addition to their activation levels, and separate weights between nodes for flow of emotional valence. As I've suggested, my guess is that emotions are not matters for coherence calculations, but are themselves manifestations of internal monitoring of coherence and incoherence.

In summary, then: suppose thinking better is better adaptability in fulfillment of basic creaturely goals. It seems a major contributor to such adaptability is a capacity for *learning*, which I propose is a feedback mechanism reporting coherence levels among cognitions and conations as an error measure. These reports of coherence levels to be used for learning are, plausibly, just the emotions.

References

Damasio, A. R. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. G. P. Putnam's Sons.

Ellsworth, P. C., and Scherer, K. R. 2003. Appraisal processes in emotion. In Davidson, R. J.; Scherer, K. R.; and Goldsmith, H. H., eds., *Handbook of Affective Sciences*. Oxford University Press. chapter 29, 572–595.

Manning, R. N. 1998. Functional explanation. In Craig, E., ed., *Routledge Encyclopedia of Philosophy*. Routledge, online edition.

Thagard, P. 1992. *Conceptual Revolutions*. Princeton University Press, 1993 edition.

Thagard, P. 2000. *Coherence in Thought and Action*. MIT Press.

---

[3]Actually, it's more like what philosophers have come to call a matter of 'foundherence", since there are defeasible foundations to the problem represented in the default nodes connected to the world. This proposal builds on work by Paul Thagard toward modeling cognitive coherence; see especially (Thagard 2000).

[4]See (Ellsworth & Scherer 2003) for a recent overview.

[5]See (Damasio 1994).

[6]My suspicion is that with an internal learning standard like mine, his DECO, ECHO *etc.* can be subsumed into one overall system. For example my learning standard may help ECHO to learn better parameter settings through feedback, a puzzle from (Thagard 1992) p81. My standard may also be able to arbitrate the connections needed *between* ECHO and DECO, and so on.