Review

Biological Psychiatry

A Formal Valuation Framework for Emotions and Their Control

Quentin J.M. Huys and Daniel Renz

ABSTRACT

Computational psychiatry aims to apply mathematical and computational techniques to help improve psychiatric care. To achieve this, the phenomena under scrutiny should be within the scope of formal methods. As emotions play an important role across many psychiatric disorders, such computational methods must encompass emotions. Here, we consider formal valuation accounts of emotions. We focus on the fact that the flexibility of emotional responses and the nature of appraisals suggest the need for a model-based valuation framework for emotions. However, resource limitations make plain model-based valuation impossible and require metareasoning strategies to apportion cognitive resources adaptively. We argue that emotions may implement such metareasoning approximations by restricting the range of behaviors and states considered. We consider the processes that guide the deployment of the approximations, discerning between innate, model-free, heuristic, and model-based controllers. A formal valuation and metareasoning framework may thus provide a principled approach to examining emotions.

Keywords: Computational psychiatry, Decision making, Emotion regulation, Emotions, Model based, Reinforcement learning

http://dx.doi.org/10.1016/j.biopsych.2017.07.003

Computational psychiatry is a young field hoping to leverage advances in computational techniques to understand and improve mental health (1–5). It is motivated on the one hand by the necessity to bring novel statistical and machine-learning techniques to bear on the rapidly expanding complexity of novel datasets relevant to mental health, and on the other hand by the complexity of the problem itself as mental health relates to the most difficult tasks performed by the most complex of organs.

Emotions are central to mental health, and emotional disorders contribute substantially to the burden of mental illnesses (6). The traditional dichotomization of emotion and reason might question the feasibility of applying computational techniques to the core issues of emotion. It is therefore imperative for computational psychiatry that we consider the ability of a computational and mathematical framework to address core emotional phenomena. Here, we argue that approaching emotion computationally requires the introduction of model-based valuation and metareasoning. Metareasoning considers optimal valuation in the face of resource constraints (7–9). The proposal is that human emotions involve strategies to deal with the complexity of model-based or goal-directed decision making by focusing on particular aspects of the problem at hand.

Research on human emotions is complicated as questions about their nature continue to divide the scientific community (10,11). Nevertheless, there is consensus on a number of key components that characterize emotions, and this review attempts to view them in a computational light. We first provide a description of important features of emotions, then introduce valuation and the metareasoning problem, then relate approximate metareasoning strategies to features of emotions, and finally describe the control of approximate metareasoning strategies.

INGREDIENTS OF A COMPUTATIONAL APPROACH TO EMOTIONS

Key features of human emotions that require accounting for and that are emphasized to various degrees in different conceptualizations are 1) correlated physiological, psychological and behavioral processes shaped by evolutionarily predefined neural circuitry; 2) interpretations or appraisals; and 3) conscious verbal self-report about emotions. Key problems in contemporary research on human emotions include to what extent the three feature domains are related (e.g., how conscious emotions in humans relate to evolutionarily predefined circuitry) and to what extent emotions are discrete entities.

Basic emotion theories suggest that there are a limited, relatively fixed, number of universal, evolutionarily shaped, culture-independent, and neurobiologically hard-coded emotional categories including happiness, surprise, sadness, disgust, anger, and fear (11–13). For the present purpose, what is important is that these represent a set of innately interlinked physiological, behavioral, and psychological processes that are triggered in an inflexible manner by species-specific salient stimuli, akin to unconditioned responses. Animal research, in which specific responses to species-relevant stimuli are

ISSN: 0006-3223

observable and readily quantifiable, has contributed to this view. However, behavioral responses in animals cannot be directly translated to emotional experiences in humans. Amygdalar and hippocampal damage, for instance, abolish physiological and autobiographical signatures of aversive conditioning, respectively, while leaving the other intact (14). Furthermore, aversive conditioning can be performed subliminally and can evoke amygdala activity and physiological response, but can fail to result in any emotion of fear (15,16), while amygdalar lesions can leave human fear unaffected (17,18).

Human emotional responses to stimuli are characterized by substantial within- and between-subject variability. Appraisal theory locates one explanation for this variability in the interpretation (be it conscious or unconscious) of a particular situation or stimulus as being relevant to the individual's goals (19). This interpretation depends on the goal and the individual's beliefs in addition to the stimulus. A stimulus or situation being interpreted as increasing the changes of reaching one's goals would, for instance, result in the emotion of joy or happiness (20–22). However, just like basic emotion theories, appraisal theories often view the expressed emotion itself as a "definable pattern of outputs that preexist within the individual" (10). For instance, Scherer (23) defined them as "episode[s] of interrelated, synchronized changes in the states of all or most […] organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism."

The evidence for discrete emotions is controversial. Autonomic responses, electroencephalographic features, and facial expressions do not permit simple categorization and show little evidence of the predicted correlations (10,24,25), though newer machine learning approaches have shown that categorical information can be extracted from physiological (26) and neural (27,28) data. The latter analyses have, however, clarified that there is no single underlying substrate for particular emotions. Rather, each emotional category depends on a distributed network of limbic but also cortical components that reflect the particular neurocognitive processes involved (29).

An alternative view is that the discreteness of emotions arises from the categorical labeling of internal events for the purpose of intra- or intersubject communication. Neuroimaging has provided some support for such a model, arguing that the ventrolateral prefrontal cortex is involved in categorical labeling of emotional states (30–32) evolving along the two major axes of valence (from good to bad) and arousal (from high to low). Indeed, factor analyses of a variety of measures of emotions including similarity ratings among words, facial expression, and autonomic measures reliably identify these two separate dimensions (33). Neuroimaging has also been used to argue that while the amygdala tracks arousal, the orbitofrontal cortex tracks valence across emotions (34).

VALUATION AND EMOTION

Basic and animal emotion research, with its grounding in evolutionarily shaped responses, emphasizes the importance of emotions in guiding behavior adaptively. A focus on adaptive responding is also present in appraisal theories, which suggest that emotions arise when events are judged to be relevant to the individual's "needs, attachments, values, current goals and beliefs" (35). Computationally, inferring adaptive choices involves integrating not only immediate rewards, but also longer-term rewards, and for that reason requires consideration of the future course of events. This evaluation of the future is where the problem lies, as the further into the future one looks, the broader the range of potential events. Specifically, valuation involves summing over an exponentially expanding decision tree of future possibilities. Optimal valuation would search the entire tree, which is rarely feasible. Reinforcement learning is a thriving subfield of machine learning concerned with algorithmic solutions to this problem.

Model-Free Accounts of Emotional Expression

A substantial body of work has related one such algorithmic solution to how emotional expressions change over time (36). In so-called model-free reinforcement learning, the stability of the world is exploited to replace integration over the future with actual past experiences. Clever bookkeeping allows the use of prediction errors to update values that, in the limit of extensive experience, are guaranteed to yield the true long-term values of states and behaviors (37). Here, emotional responses are viewed as a type of high-level action, involving multiple biological and neural subsystems. One example of such an "action" is a freezing response, which has behavioral, attentional, and physiological components. These high-level actions are thought to be emitted either in an innate fashion (38) in response to the appropriate species-specific unconditional stimulus (39–41), or after learning in response to a conditioned stimulus. In the latter case, the expression of the action is proportional to the value attached to the conditioned stimulus, which in turn is a scalar measure of the average expected unconditional stimulus strength (42–44). This has been applied to a wide variety of affective responses, including heart rate changes (45), approach (46), avoidance (47,48), extinction (49), vigor (50,51), and others. Perhaps the most striking success of these models is their ability to capture how pavlovian affective responses can lead to maladaptive choices (43,52).

Model-free approaches are very valuable to understand how the expression of affect transfers between situations with experience. Although mostly restricted to individual laboratory sessions, the underlying model likely plays an important role in explaining how individual differences in the expression of (affective) behaviors emerge over (life)time, and potentially in response to behavioral psychotherapeutic interventions. Furthermore, a hierarchical version of model-free reinforcement learning has the capacity to explain how complex high-level actions consisting of multiple correlated processes might emerge (53–55), though this awaits application to the correlations among physiological, psychological, and behavioral aspects of emotions.

Appraisals Require Model-Based Inference

Pure model-free accounts, however, fail to explain context effects on conditioning. For instance, the physiological response to a threat differs depending on whether the animal is restrained or freely moving (56) as well as whether a refuge or obstacle is present and at what distance (57–59). In humans, framing the same movement as approach or withdrawal alters whether a pavlovian conditioned stimulus promotes or inhibits