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

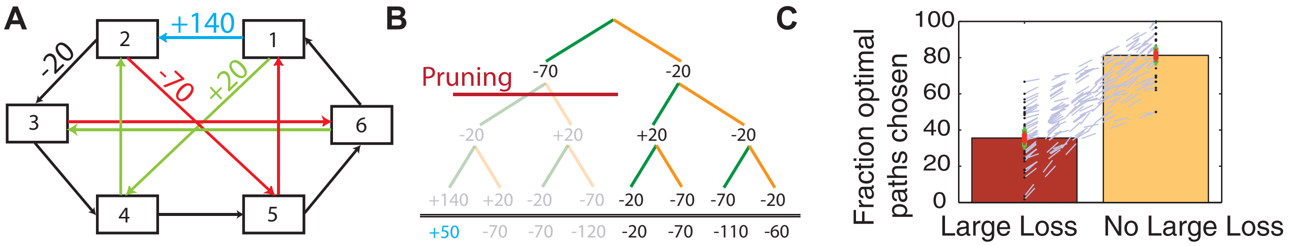

**Figure 1.** Model-based inference problems rapidly become too hard to fully solve. Consider a task **(A)** where subjects see six rectangles arranged hexagonally and are taught to navigate according to the underlying transition matrix indicated by the arrows, so that they can transition from each state to two successor states depending on which of two buttons they press. Every transition yields a reward or loss. Finding a path of a given length that maximizes total earnings corresponds to a tree-shaped decision problem, for instance the one in panel **(B)** for three transitions starting from state 3. Participants typically choose not to expend cognitive effort evaluating subtrees below a salient loss (here below transitions with –70 points), resulting in a cutting of or "pruning" of decision trees **(B)**. **(C)** This in turn results in worse performance when the optimal path requires transitioning through a salient loss (red) than when it does not (orange). Black dots and gray lines show the effect in individual participants. **(A)** and **(B)** are modified with permission from Huys *et al.* (44). **(C)** shows data replotted from Huys *et al.* (55).

it (42). Even startle reflexes are potentiated by fear induction (60), enhanced by upregulating negative emotions (61), and reduced by positive emotions (62). More fundamentally, context determines what affective behavior is emitted (21,63): the same emotion of anger may motivate harm not only through physical means in a boxing ring, but also through financial transactions in a boardroom setting.

Similarly, scalar summaries of past experience cannot account for the impact of appraisals on human emotions. The appraisal of an event involves an assessment of what the event "means" (35). This interpretative process determines whether and which emotion results by inference of latent causes: a smile is pleasant if interpreted as emanating from kindness, but aversive if viewed as an expression of condescension. The fact that an event is meaningful can be inferred from changes in model-free values because these capture the expectation of future well-being in a relatively stable environment, and sudden changes in model-free expected values therefore indicate a meaningful event. This is reminiscent of the argument that changes in core affect invoke appraisals (10,33). Certain aspects of meaning may also be precomputed and result in automatic appraisals (64,65), but *what* an event means for well-being cannot be derived from model-free values. This assessment involves a series of variables such as goal congruence, controllability, and agency (35,66,67) that capture how the changed contingencies induced by the event and the behavior influence the controllable achievability of the goal (68). Goal congruence, for instance, measures how events influence the ability and cost of achieving current goals and as such involves replanning a new path toward the goal and comparing the cost of this path to that of the previous plan.

The computation of the meaning part of the appraisal requires the integration of a model capturing an individual's beliefs about the consequences of choices, what reinforcements will be obtained in which states, how observations relate to hidden states, and how different states relate to each other (37,65,69). Inferring such values requires a model to be inverted or simulated. Mirroring the notion that some appraisals rely on rule-based processes, it suggests a role for nonautomatic components; it captures that appraisals generalize and change over time as new information is progressively integrated; it suggests how maladaptive beliefs influence emotions; and it suggests how a new understanding can alter emotions in explicit reappraisal (35,63,66,70,71).

However, a measured and "rational" consideration of all possible outcomes is hardly a sufficient model of emotions (72). In fact, reasoning itself is profoundly affected by emotions (70), as are perception, learning, and memory.

## METAREASONING

One factor that may be useful to consider is that model-based inference is mostly impossible due to the sheer size of most relevant model-based valuation problems. Figure 1A shows a simple planning task in a maze the solution of which corresponds to a binary decision tree (Figure 1B). The best action at the root of the tree is the one that leads to the path with the best total outcome, and this may not be the action with the best immediate reward. As in reality, there is usually more than one way to the goal, but the different paths have different intermediate outcomes rendering some better than others. Despite its simplicity, humans have difficulty solving the task fully, and employ strategies to avoid evaluating the entire tree even when there are only three or four choices to go (Figure 1C). Unless they are highly constrained, such as in feedforward motor control (73), optimal decisions in realistic situations are computationally extremely demanding.

The limited resources lead to the metareasoning problem, which concerns the optimal deployment of the available computational power (7,8,74,75). It is a decision problem about which of the various options to evaluate internally (Figure 2). Formally, the estimated value of performing a computation is the difference in expected utility between taking a choice without the additional computation, and taking a new alternative choice after having invested in the computation (8,76). Although this decision problem is mathematically similar to the original problem, it is different from the original problem because simulations do not actually incur the costs of the real problem, and while taking real poor actions should be avoided to avoid incurring their loss, internally simulating poor actions can be useful (77,78). Thus, the states in this metareasoning problem are all possible partial trees of the original tree, which is a far larger state space than in the original problem.

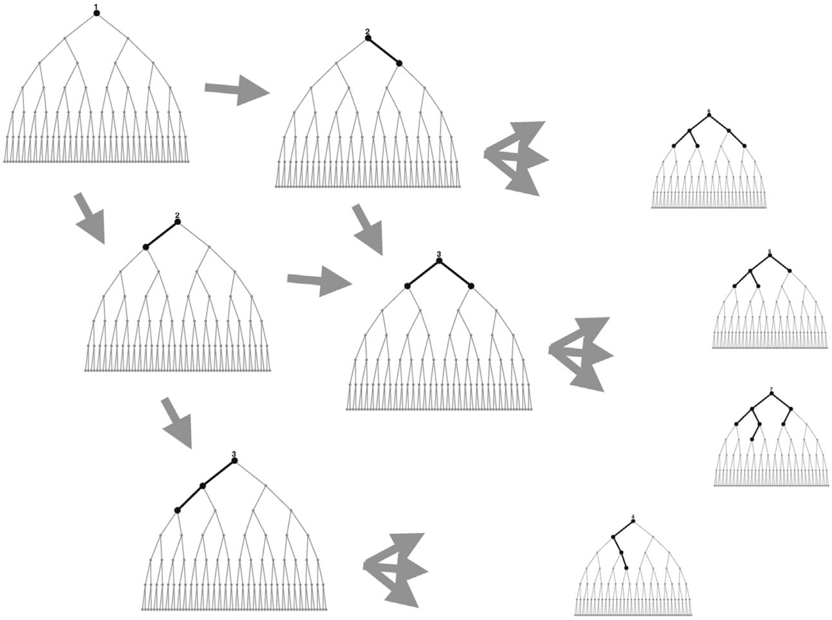

**Figure 2.** Metareasoning. Given a state and a sequence of possible decisions, optimal action choice involves evaluating a decision tree (top left). An agent with constrained resources faces the challenge of deciding what to simulate, i.e., what to "think about." For instance, it could choose to first simulate the action going left, and then continue down this branch (leftmost set of trees). Alternatively, it could superficially consider the right action, and then start examining the left action. The metareasoning problem hence is a decision problem where the states are knowledge states about the decision tree, and the task is to choose to think about the components of the problem at hand in the way that is most likely to yield a good final choice.

## Emotions Implement Approximate Metareasoning Strategies

Model-based reasoning is hence faced with two profound challenges: the size of the problem and the even harder task of apportioning limited resources in an adaptive manner. These are fundamental problems and a strategy to deal with them is mandatory. The proposal here is that emotions can implement approximate solutions to these challenges. In particular, emotional states 1) come with a strong focus on particular behaviors and 2) induce a strong perceptual and processing focus such that evaluation is concentrated on a narrow set of states. Emotions thereby effectively function as approximate metareasoning strategies that prescribe how computational resources are allocated. To what extent these approximations are adaptive depends on how they are invoked. We first provide an outline for the implementation in terms of action tendencies and state observations, and then turn to the control issues.

**Action Tendencies.** One of the features of emotions about which there is more agreement is that they prioritize certain actions (13,35,79–81). Constraining the action space can substantially simplify the valuation problem because the computational cost is exponential in the size of the action space.

At an abstract level, emotional states are accompanied by distinct and richly experienced urges toward particular classes of actions. Frijda *et al.* (20) asked people to remember events of particular emotions and then to rate a list or 26 items about the kinds of behaviors they wanted to engage in, such as "I gave up," "I wanted to protect myself from someone or something," or "I wanted to help someone, to take care of someone." From the ratings of these statements, the emotion characterizing the episode could be reliably recovered. Though very abstract, such rich descriptions are also important in psychotherapeutic settings. In dialectic behavior therapy, individuals are initially taught to recognize emotions by the action tendencies they feel (71).

Emotions also induce physiological and vegetative changes. However, physiological signatures of emotions do not appear to readily differentiate between categorically defined emotions, but rather provide a few classes of general action preparations [(79,82–84), though see (26)]. A preparatory increase in heart rate to compensate for the anticipated drop in peripheral resistance upon supplying blood to large muscle groups is required when running, be it for escape or fun. As such, these can be seen as a preparation toward a class of behaviors that share physiological requirements.

**State Observation.** The complexity of model-based evaluation is also exponential in the range of states considered. There is ample evidence for emotion- or mood-congruent processing biases (85,86). For instance, Bradley *et al.* (87) showed that exposure to sad music and recollection of sad memories produces an attentional bias toward sad words, and such biases arise from the emotional state rather than purely from the exposure to the emotional word (88). By restricting attention to particular states and disregarding others, the problem could again be reduced in size (89), for instance by pruning (44) searches along branches of the decision tree that result in states outside the attentional focus.

A further aspect is that there is usually uncertainty about the state. This profoundly complicates the computational task of valuation because policies for the various possible states have to be computed (90). By ascertaining the state, this complexity can be reduced. Introspection about the state of the body likely plays a particularly important role: the impact of a muscle's activation depends on the joint position and the chances of success in a fight are reduced when already wounded.

## Controlling Metareasoning Strategies

If there are multiple approximate metareasoning strategies, then there must be some control over which is deployed when. The first source of control is likely evolutionary, where species-specific responses provide a (potentially very strong) bias toward evaluating particular actions, rather than toward emitting the action entirely inflexibly. This allows for the kinds of context effects on even innate responses mentioned above.

The second source of control could be, confusingly, model-free. Performers may learn from experience that a certain amount of catastrophizing improves their performance (91), without an understanding of why that is. Etkin *et al.* (92) have recently argued for a model-free component in serial adaptations in the emotional conflict task, where individuals have to indicate the facial expression (fear or happy) of a face with either the matching or conflicting word superimposed over it (93). Model-free learning has been argued to account for learning in strategy selection (94): with repeated experience, individuals can slowly increase their frequency of using adaptive strategies for solving problems (95). We have recently shown how the results of costly model-based evaluations are memorized and simply replayed upon repeated encounter of the same problem in a process called memoization (55) that gives rise to decision-making biases that are characteristic of the individual but highly variable across the population.

The third evaluative process for emotions allows for knowledge to be incorporated in the form of heuristics. Research on decisions about options with many attributes [for instance cars, with price, speed, size, brand, etc. (96)] have identified a host of different strategies. “Take the best” is appropriate in noncompensatory environments where one feature is most informative and can be used alone to rank options. In compensatory environments, humans spend more time and cognitive effort on examining multiple features and integrating the information, but only if they are not under time pressure (97). This suggests that individuals can access approximate measures of how adaptive a particular cognitive strategy is, and use this to guide their choice (98). In the affective domain, misguided beliefs or schemas (99) about the adaptiveness of strategies relate to a number of pathological emotion regulation phenomena. Pathological worry in generalized anxiety disorders (100,101) and rumination in depressive disorders (102) are maintained by explicit beliefs about the usefulness of worry and rumination, respectively. People who dislike emotion regulation are more likely to respond with anger to provocation (103). Depressed persons are not impaired at emotion regulation strategies such as positive imagery to improve their mood, but they have a reduced tendency to employ them (104).

The fourth evaluative process, again confusingly, could be model-based, where the precise consequences of particular emotions are examined and evaluated. This is rarely feasible and probably only commonly done in situational analyses in psychotherapy, where emotions, thoughts, behavior, and consequences are explicitly discussed (71,99,105). This allows patients to learn to consciously and explicitly assess whether a particular emotion is appropriate and helpful in a given situation, and to adapt it by using reappraisal and other emotion regulation strategies if necessary.

## DISCUSSION

We have attempted to sketch out a valuation framework for emotions. We have argued why appraisals point toward a model-based framework; how emotions may have a potentially important role in facilitating model-based decisions by functioning as internal strategies to allocate computational resources; how emotions’ adaptive nature depends on their deployment; and how a variety of different processes can lead to adaptive or maladaptive deployment of emotion strategies.

Three desiderata for a computational framework of emotions were put forth. The first was the at best partially correlated nature of physiological, psychological, and behavioral features. The flexibility the proposed framework allows for contrasts with the view of basic emotions as relatively fixed behavioral and physiological action packages. As such, it reflects the lack of identifiably discrete physiological or behavioral patterns or single neurobiological cause (10,24,29). Similar to other proposals, it emphasizes the importance of emotional processes in more complex decision-making settings (106). Space constraints have prevented us from exploring the distinction between valence and arousal important to circumplex and core affect models, but this might naturally emerge from valuation in continuous-time settings, where the rate at which actions are emitted depends on the average reward rate in the environment, albeit in sometimes complex ways (36,50,51,107). Notably, the current proposal allows for mixed emotions through a combination of metareasoning policies.

The second desideratum was the ability to account for appraisal and contextual effects. The complexity of the model-based valuation required for this led to the notion of approximate metareasoning strategies. These approximate strategies are necessarily often suboptimal and may capture the prototypical adverse influences of emotion on cognition (108). The focus on valuation is compatible with models emphasizing prediction (109), but distinct in that it suggests that the relevant predictions must be about long-term utility, and that emotions play a key role in facilitating such predictions, albeit approximately.

The third desideratum concerned the nature of conscious qualia of emotions. Doing so fully awaits a theory of consciousness. However, two aspects are interesting to consider. First, Dehaene and Naccache’s (110) notion of a global workspace sits naturally with the notion of metareasoning. Situations with a high estimated value of computation should recruit neural resources more extensively, and hence be more likely to involve the brain-wide states postulated as representing the global workspace. Second, it has been suggested that the component processes in verbal self-report involve an interoceptive component followed by a classification process (30). In our proposal, the metareasoning strategies would profoundly influence what information was processed, and as such may strongly determine future classification. Interestingly, the classification process has been suggested to involve the ventrolateral prefrontal cortex (31,32), which is also known to mediate arbitration between valuation strategies (111).

The framework laid out here makes a number of testable predictions. First, the argument that appraisal involves model-based reasoning means that it should be influenced by cognitive, endocrine, and neuromodulatory variables known to influence model-based reasoning (112–115). Second, we emphasized the importance of the control of metareasoning strategies and suggested that it may be subject to substantial malleability. This in turn predicts that by training particular metareasoning strategies it should be possible to selectively facilitate certain emotions over others. Third, it suggests that emotions do not represent valuations themselves, but rather that they determine the process by which valuation occurs. This predicts an influence of emotion on search in a relevant model-based task rather than on values directly. The main challenge for the framework is that it points to the critical importance of understanding both the internal models of individuals and their strategies in searching them. Measuring these is a difficult scientific problem. Though this is as yet not feasible, recent advances including whole-brain mapping of semantic representations (116,117) combined with active and passive sensing using mobile devices (118,119) should open promising avenues.

The emphasis on model-based processes was partially motivated by the finding that model-free measures of reward processing and learning are unimpaired in depression (120,121), and as such this work is an effort to start integrating cognitive phenomena of clinical importance like dysfunctional attitudes (122), helplessness (68), attributional/cognitive styles (123,124), and appraisals (35) into a valuation framework.

## ACKNOWLEDGMENTS AND DISCLOSURES

This work was supported by Swiss National Science Foundation Grant No. 320030L_153449/1 (to QJMH).

We thank Peter Dayan for comments on an earlier version of this manuscript.

The authors report no biomedical financial interests or potential conflicts of interest.

## ARTICLE INFORMATION

From the Translational Neuromodeling Unit (QJMH, DR), Institute for Biomedical Engineering, University of Zurich and Swiss Federal Institute of Technology (ETH Zurich); and the Centre for Addictive Disorders, Department of Psychiatry, Psychotherapy and Psychosomatics (QJMH), Hospital of Psychiatry, University of Zurich, Zürich, Switzerland.

Address correspondence to Quentin J.M. Huys, Translational Neuromodeling Unit, Wilfriedstrasse 6, 8032 Zürich, Switzerland; E-mail: qhuys@cantab.net.

Received Nov 5, 2016; revised Jun 30, 2017; accepted Jul 1, 2017.

## REFERENCES

1. Huys QJM, Moutoussis M, Williams J (2011): Are computational models of any use to psychiatry? Neural Netw 24:544–551.
2. Maia TV, Frank MJ (2011): From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci 14:154–162.
3. Montague PR, Dolan RJ, Friston KJ, Dayan P (2012): Computational psychiatry. Trends Cogn Sci 16:72–80.
4. Stephan KE, Mathys C (2014): Computational approaches to psychiatry. Curr Opin Neurobiol 25:85–92.
5. Huys QJM, Maia TV, Frank MJ (2016): Computational psychiatry as a bridge from neuroscience to clinical applications. Nat Neurosci 19:404–413.
6. Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, *et al.* (2013): Global burden of disease attributable to mental and substance use disorders: Findings from the global burden of disease study 2010. Lancet 382:1575–1586.
7. Simon HA (1956): Rational choice and the structure of the environment. Psychol Rev 63:129–138.
8. Russell S, Wefald E (1991): Principles of metareasoning. Artif Intell 49:361–395.
9. Payne JW, Bettman JR, Johnson EJ (1993): The Adaptive Decision Maker. New York: Cambridge University Press.
10. Barrett LF (2006): Are emotions natural kinds? Perspect Psychol Sci 1:28–58.
11. Panksepp J (2007): Neurologizing the psychology of affects: How appraisal-based constructivism and basic emotion theory can coexist. Perspect Psychol Sci 2:281–296.
12. Ekman P (1992): An argument for basic emotions. Cogn Emot 6:169–200.
13. Izard CE (2007): Basic emotions, natural kinds, emotion schemas, and a new paradigm. Perspect Psychol Sci 2:260–280.
14. LeDoux JE (2014): Coming to terms with fear. Proc Natl Acad Sci U S A 111:2871–2878.
15. Raio CM, Carmel D, Carrasco M, Phelps EA (2012): Nonconscious fear is quickly acquired but swiftly forgotten. Curr Biol 22:R477–R479.
16. Knight DC, Waters NS, Bandettini PA (2009): Neural substrates of explicit and implicit fear memory. Neuroimage 45:208–214.
17. Feinstein JS, Buzza C, Hurlemann R, Follmer RL, Dahdaleh NS, Coryell WH, *et al.* (2013): Fear and panic in humans with bilateral amygdala damage. Nat Neurosci 16:270–272.
18. Khalsa SS, Feinstein JS, Li W, Feusner JD, Adolphs R, Hurlemann R (2016): Panic anxiety in humans with bilateral amygdala lesions: Pharmacological induction via cardiorespiratory interoceptive pathways. J Neurosci 36:3559–3566.
19. Roseman IJ, Smith CA (2001): Appraisal Theory. Oxford, UK: Oxford University Press.
20. Frijda NH, Kuipers P, Ter Schure E (1989): Relations among emotion, appraisal, and emotional action readiness. J Pers Soc Psychol 57:212–228.
21. Lazarus RS: Stress and Emotion: A New Synthesis. New York: Springer.
22. Ellsworth PC (2013): Appraisal theory: Old and new questions. Emot Rev 5:125–131.
23. Scherer KR (2005): What are emotions? And how can they be measured? Soc Sci Information 44:695–729.
24. Cacioppo JT, Berntson GG, Larsen JT, Poehlmann KM, Ito TA, *et al.* (2000): The psychophysiology of emotion. In: Lewis R, Haviland-Jones JM, editors. Handbook of Emotions, 2nd ed. New York: Guilford Press, 173–191.
25. Jack RE, Garrod OGB, Yu H, Caldara R, Schyns PG (2012): Facial expressions of emotion are not culturally universal. Proc Natl Acad Sci U S A 109:7241–7244.
26. Stephens CL, Christie IC, Friedman BH (2010): Autonomic specificity of basic emotions: Evidence from pattern classification and cluster analysis. Biol Psychol 84:463–473.
27. Kragel PA, LaBar KS (2015): Multivariate neural biomarkers of emotional states are categorically distinct. Soc Cogn Affect Neurosci 10:1437–1448.
28. Saarimki H, Gotsopoulos A, Jskelinen IP, Lampinen J, Vuilleumier P, Hari R, *et al.* (2016): Discrete neural signatures of basic emotions. Cereb Cortex 26:2563–2573.
29. Wager TD, Kang J, Johnson TD, Nichols TE, Satpute AB, Barrett LF (2015): A Bayesian model of category-specific emotional brain responses. PLoS Comput Biol 11:e1004066.
30. Barrett LF (2006): Solving the emotion paradox: Categorization and the experience of emotion. Pers Soc Psychol Rev 10:20–46.
31. Lieberman MD, Eisenberger NI, Crockett MJ, Tom SM, Pfeifer JH, Way BM (2007): Putting feelings into words: Affect labeling disrupts amygdala activity in response to affective stimuli. Psychol Sci 18:421–428.
32. Satpute AB, Shu J, Weber J, Roy M, Ochsner KN (2013): The functional neural architecture of self-reports of affective experience. Biol Psychiatry 73:631–638.

33. Russell JA (2003): Core affect and the psychological construction of emotion. Psychol Rev 110:145–172.
34. Wilson-Mendenhall CD, Barrett LF, Barsalou LW (2013): Neural evidence that human emotions share core affective properties. Psychol Sci 24:947–956.
35. Moors A, Ellsworth PC, Scherer KR, Frijda NH (2013): Appraisal theories of emotion: State of the art and future development. Emot Rev 5:119–124.
36. Bach DR, Dayan P (2017): Algorithms for survival: A comparative perspective on emotions. Nat Rev Neurosci 18:311–319.
37. Sutton RS, Barto AG (1998): Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press.
38. Hirsch S, Bolles R (1980): On the ability of prey to recognize predators. Z Tierpsychol 54:71–84.
39. Bolles RC (1970): Species-specific defense reactions and avoidance learning. Psychol Rev 77:32–48.
40. Seligman ME (1970): On the generality of the laws of learning. Psychol Rev 77:406–418.
41. Timberlake W, Wahl G, King D (1982): Stimulus and response contingencies in the misbehavior of rats. J Exp Psychol Anim Behav Process 8:62–85.
42. Huys QJM, Cools R, Gölzer M, Friedel E, Heinz A, Dolan RJ, *et al.* (2011): Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. PLoS Comput Biol 7:e1002028.
43. Dayan P, Niv Y, Seymour B, Daw ND (2006): The misbehavior of value and the discipline of the will. Neural Netw 19:1153–1160.
44. Huys QJM, Eshel N, O'Nions E, Sheridan L, Dayan P, Roiser JP (2012): Bonsai trees in your head: How the Pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput Biol 8:e1002410.
45. Seymour B, O'Doherty JP, Koltzenburg M, Wiech K, Frackowiak R, Friston K, *et al.* (2005): Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. Nat Neurosci 8:1234–1240.
46. Lesaint F, Sigaud O, Flagel SB, Robinson TE, Khamassi M (2014): Modelling individual differences in the form of pavlovian conditioned approach responses: A dual learning systems approach with factored representations. PLoS Comput Biol 10:e1003466.
47. Moutoussis M, Bentall RP, Williams J, Dayan P (2008): A temporal difference account of avoidance learning. Network 19:137–160.
48. Maia TV (2010): Two-factor theory, the actor-critic model, and conditioned avoidance. Learn Behav 38:50–67.
49. Gershman SJ, Blei DM, Niv Y (2010): Context, learning, and extinction. Psychol Rev 117:197–209.
50. Niv Y, Daw ND, Joel D, Dayan P (2007): Tonic dopamine: Opportunity costs and the control of response vigor. Psychopharmacology (Berl) 191:507–520.
51. Dayan P (2012): Instrumental vigour in punishment and reward. Eur J Neurosci 35:1152–1168.
52. Breland K, Breland M (1961): The misbehavior of organisms. Am Psychol 16:681–684.
53. Sutton RS, Precup D, Singh S (1999): Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artif Intell 112:181–211.
54. Botvinick MM, Niv Y, Barto AC (2009): Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. Cognition 113:262–280.
55. Huys QJM, Lally N, Faulkner P, Eshel N, Seifritz E, Gershman SJ, *et al.* (2015): Interplay of approximate planning strategies. Proc Natl Acad Sci U S A 112:3098–3103.
56. Iwata J, LeDoux JE (1988): Dissociation of associative and non-associative concomitants of classical fear conditioning in the freely behaving rat. Behav Neurosci 102:66–76.
57. Domenici P, Blagburn JM, Bacon JP (2011): Animal escapology II: Escape trajectory case studies. J Exp Biol 214:2474–2494.
58. Domenici P, Blagburn JM, Bacon JP (2011): Animal escapology I: Theoretical issues and emerging trends in escape trajectories. J Exp Biol 214:2463–2473.
59. Zani P, Jones T, Neuhaus R, Milgrom J (2009): Effect of refuge distance on escape behavior of side-blotched lizards (uta stansburiana). Can J Zoology 87:407–414.
60. Grillon C, Davis M (1997): Fear-potentiated startle conditioning in humans: explicit and contextual cue conditioning following paired versus unpaired training. Psychophysiology 34:451–458.
61. Jackson DC, Malmstadt JR, Larson CL, Davidson RJ (2000): Suppression and enhancement of emotional responses to unpleasant pictures. Psychophysiology 37:515–522.
62. Lang PJ, Bradley MM, Cuthbert BN (1990): Emotion, attention, and the startle reflex. Psychol Rev 97:377–395.
63. Lazarus RS, Alfert E (1964): Short-circuiting of threat by experimentally altering cognitive appraisal. J Abnorm Psychol 69:195–205.
64. Moors A (2010): Automatic constructive appraisal as a candidate cause of emotion. Emot Rev 2:139–156.
65. Daw ND, Dayan P (2014): The algorithmic anatomy of model-based evaluation. Philos Trans R Soc Lond B Biol Sci 369:20130478.
66. Lazarus RS (1994): Emotion and Adaptation. Oxford, UK: Oxford University Press.
67. Scherer KR (1984): On the nature and function of emotion: A component process approach. In: Scherer KR, Ekman P, editors. Approaches to Emotion. Hillsdale, NJ: Erlbaum, 293–317.
68. Huys QJM, Dayan P (2009): A Bayesian formulation of behavioral control. Cognition 113:314–328.
69. Daw ND, Niv Y, Dayan P (2005): Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8:1704–1711.
70. Beck AT (1967): Depression: Clinical, Experimental and Theoretical Aspects. New York: Harper & Row.
71. Bohus M, Wolf-Arehult M (2013): Interactive Skill Training for Borderline Patients [German]. Stuttgart, Germany: Schattauer Verlag.
72. Zajonc RB (1980): Feeling and thinking: Preferences need no inferences. Am Psychol 35:151.
73. Kawato M (1999): Internal models for motor control and trajectory planning. Curr Opin Neurobiol 9:718–727.
74. Payne JW, Bettman JR, Luce MF (1996): When time is money: Decision behavior under opportunity-cost time pressure. Organ Behav Hum Dec Process 66:131–152.
75. Lieder F, Griffiths TL, Huys QJ, Goodman ND (2017): The anchoring bias reflects rational use of cognitive resources [published online ahead of print May 8]. Psychon Bull Rev.
76. Keramati M, Dezfouli A, Piray P (2011): Speed/accuracy trade-off between the habitual and the goal-directed processes. PLoS Comput Biol 7:e1002055.
77. Bubeck S, Munos R, Stoltz G (2011): Pure exploration in finitely-armed and continuous-armed bandits. Theor Comput Sci 412:1832–1852.
78. Hay N, Russell SJ (2011): Metareasoning for Monte Carlo tree search. Technical Report, Electrical Engineering and Computer Sciences. Berkeley, CA: University of California at Berkeley.
79. James W (1980): The Principles of Psychology. New York: Dover.
80. Posner J, Russell JA, Peterson BS (2005): The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev Psychopathol 17:715–734.
81. Lang PJ, Bradley MM (2013): Appetitive and defensive motivation: Goal-directed or goal-determined? Emot Rev 5:230–234.
82. Barrett LF, Ochsner KN, Gross JJ (2007): On the automaticity of emotion. In: Bargh JA, editor. Social Psychology and the Unconscious: The Automaticity of Higher Mental Processes. Suffolk, UK: Psychology Press, 173–217.
83. Friedman BH (2010): Feelings and the body: The Jamesian perspective on autonomic specificity of emotion. Biol Psychol 84:383–393.
84. Lang PJ, Bradley MM (2010): Emotion and the motivational brain. Biol Psychol 84:437–450.
85. MacLeod C, Mathews A, Tata P (1986): Attentional bias in emotional disorders. J Abnorm Psychol 95:15–20.
86. Mathews A, MacLeod C (2005): Cognitive vulnerability to emotional disorders. Annu Rev Clin Psychol 1:167–195.

87. Bradley BP, Mogg K, Lee SC (1997): Attentional biases for negative information in induced and naturally occurring dysphoria. Behav Res Ther 35:911–927.
88. Gilboa-Schechtman E, Revelle W, Gotlib IH (2000): Stroop interference following mood induction: Emotionality, mood congruence, and concern relevance. Cogn Therapy Res 24:491–502.
89. Dean T, Kaelbling LP, Kirman J, Nicholson A (1995): Planning under time constraints in stochastic domains. Artif Intell 76:35–74.
90. Kaelbling LP, Littman ML, Cassandra AR (1998): Planning and acting in partially observable stochastic domains. Artif Intell 101:99–134.
91. Wolfe ML (1989): Correlates of adaptive and maladaptive musical performance anxiety. Med Prob Perform Artists 4:49–56.
92. Etkin A, Büchel C, Gross JJ (2015): The neural bases of emotion regulation. Nat Rev Neurosci 16:693–700.
93. Etkin A, Egner T, Peraza DM, Kandel ER, Hirsch J (2006): Resolving emotional conflict: A role for the rostral anterior cingulate cortex in modulating activity in the amygdala. Neuron 51:871–882.
94. Marewski JN, Link D (2014): Strategy selection: An introduction to the modeling challenge. Wiley Interdiscip Rev Cogn Sci 5:39–59.
95. Rieskamp J, Otto PE (2006): SSL: A theory of how people learn to select strategies. J Exp Psychol Gen 135:207–236.
96. Gigerenzer G, Goldstein DG (1996): Reasoning the fast and frugal way: Models of bounded rationality. Psychol Rev 103:650–669.
97. Newell BR, Shanks DR (2003): Take the best or look at the rest? Factors influencing “one-reason” decision making. J Exp Psychol Learn Mem Cogn 29:53–65.
98. Lieder F, Plunkett D, Hamrick JB, Russell SJ, Hay N, Griffiths T (2014): Algorithm selection by rational metareasoning as a model of human strategy selection. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, editors. Advances in Neural Information Processing Systems 27: 28th Annual Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates, 2870–2878.
99. Beck AT (1987): Cognitive models of depression. J Cog Psychotherapy Int Quart 1:5–37.
100. Wells A (1999): A metacognitive model and therapy for generalized anxiety disorder. Clin Psychol Psychother 6:86–95.
101. Borkovec T, Ray WJ, Stober J (1998): Worry: A cognitive phenomenon intimately linked to affective, physiological, and interpersonal behavioral processes. Cogn Therapy Res 22:561–576.
102. Treynor W, Gonzalez R, Nolen-Hoeksema S (2003): Rumination reconsidered: A psychometric analysis. Cogn Therapy Res 27:247–259.
103. Mauss IB, Evers C, Wilhelm FH, Gross JJ (2006): How to bite your tongue without blowing your top: Implicit evaluation of emotion regulation predicts affective responding to anger provocation. Pers Soc Psychol Bull 32:589–602.
104. Ehring T, Tuschen-Caffier B, Schnlle J, Fischer S, Gross JJ (2010): Emotion regulation and vulnerability to depression: Spontaneous versus instructed use of emotion suppression and reappraisal. Emotion 10:563–572.
105. Gross JJ (1998): Antecedent- and response-focused emotion regulation: Divergent consequences for experience, expression, and physiology. J Pers Soc Psychol 74:224–237.
106. Forgas JP (1995): Mood and judgment: The affect infusion model (aim). Psychol Bull 117:39–66.
107. Constantino SM, Daw ND (2015): Learning the opportunity cost of time in a patch-foraging task. Cogn Affect Behav Neurosci 15:837–853.
108. Loewenstein G (1996): Out of control: Visceral influences on behavior. Org Behav Hum Decis Process 65:272–292.
109. Barrett LF (2017): The theory of constructed emotion: An active inference account of interoception and categorization. Soc Cogn Affect Neurosci 12:1–23.
110. Dehaene S, Naccache L (2001): Toward a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. Cognition 79:1–37.
111. Lee SW, Shimojo S, O’Doherty JP (2014): Neural computations underlying arbitration between model-based and model-free learning. Neuron 81:687–699.
112. Otto AR, Gershman SJ, Markman AB, Daw ND (2013): The curse of planning: Dissecting multiple reinforcement-learning systems by taxing the central executive. Psychol Sci 24:751–761.
113. Otto AR, Raio CM, Chiang A, Phelps EA, Daw ND (2013): Working-memory capacity protects model-based learning from stress. Proc Natl Acad Sci U S A 110:20941–20946.
114. Deserno L, Huys QJM, Boehme R, Buchert R, Heinze HJ, Grace AA, *et al.* (2015): Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. Proc Natl Acad Sci U S A 112:1595–1600.
115. Wunderlich K, Smittenaar P, Dolan RJ (2012): Dopamine enhances model-based over model-free choice behavior. Neuron 75:418–424.
116. Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016): Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532:453–458.
117. Huth AG, Lee T, Nishimoto S, Bilenko NY, Vu AT, Gallant JL (2016): Decoding the semantic content of natural movies from human brain activity. Front Syst Neurosci 10:81.
118. Torous J, Kiang MV, Lorme J, Onnela JP (2016): New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 3:e16.
119. Baker J, Mueller N, Onnela JP, Ongur D, Buckner R (2017): 368-deep dynamic phenotyping: Neural changes underlying fluctuations in bipolar disorder over one year. Biol Psychiatry 81(suppl 15): S150–S151.
120. Huys QJM, Dayan P, Daw. (2015): Depression: A decision-theoretic account. Annu Rev Neurosci 38:1–23.
121. Rutledge RB, Moutoussis M, Smittenaar P, Zeidman P, Taylor T, Hrynkiewicz L, *et al.* (2017): Association of neural and emotional impacts of reward prediction errors with major depression. JAMA Psychiatry 74:790–797.
122. Weissman A, Beck A (1978): Development and validation of the Dysfunctional Attitude Scale. Presented at the Annual Meeting of the Association for the Advancement of Behavior Therapy, Chicago.
123. Peterson C, Semmel A, von Baeyer C, Abramson L, Metalsky G, Seligman M (1982): The attributional Style Questionnaire. Cogn Therapy Res 6:287–299.
124. Abramson LY, Alloy LB, Hogan ME, Whitehouse WG, Cornette M, Akhavan S, *et al.* (1998): Suicidality and cognitive vulnerability to depression among college students: A prospective study. J Adolesc 21:473–487.