3

The Three-Process Model of Implicit and Explicit Emotion

Ryan Smith

Introduction

A central feature of Lane, Ryan, Nadel, and Greenberg's (2015) memory reconsolidation paper was the integrated memory model (IMM), which specified that whenever episodic memory, semantic memory, or emotional responses are activated, the other two are activated as well. Space did not permit detailed elaboration of what was meant by "emotional responses" in this context, but a key concept within the paper was the distinction between implicit and explicit processes. Since these processes play an important role in psychopathology, and its treatment with psychotherapy, a more in-depth discussion of this topic is needed.

The primary aims of the present chapter are (a) to provide an overview of a range of emotion-related phenomena that have been referred to as "implicit" or "unconscious"; (b) to review a previously proposed neuro-cognitive model of conscious and unconscious emotion—hereafter referred to as the "three-process model" (R. Smith, Killgore, & Lane, 2017)—that organizes and accounts for the aforementioned phenomena; and (c) to highlight the relevance of this model to clinical psychology and psychiatry (e.g., the process of change in psychotherapy). With respect to the third aim, I will specifically illustrate how the three-process model can provide mechanistic explanations regarding:

1. Individual differences in trait emotional awareness (tEA) and their role in psychopathology and treatment.
2. The role of certain types of past experience and learned expectations in facilitating perceptions, thoughts, and actions that can promote/maintain emotional pathology.

Ryan Smith, *The Three-Process Model of Implicit and Explicit Emotion* In: *Neuroscience of Enduring Change*. Edited by: Richard D. Lane and Lynn Nadel, Oxford University Press (2020).
DOI: 10.1093/oso/9780190881511.003.0003

3. How the perceptual and cognitive causes and/or consequences of emotional responses can remain unconscious (and/or poorly understood), and how this could contribute to emotional pathology.
4. How various components of empirically supported psychotherapeutic treatments—such as reappraisal, mindfulness/acceptance, and exposure—can be understood to intervene on different neuro-cognitive processes in the model.

Before addressing these primary aims directly, however, it will also be important to first review some foundational, domain-general theories within cognitive and computational neuroscience. These domain-general theories provide the larger context within which the three-process model was developed to account for the aforementioned emotion-related phenomena.

Therefore, the chapter will be organized as follows. In the first section, I will briefly review empirical findings pertaining to three different categories of implicit/unconscious emotion. In the second section, I will then review select elements of some domain-general theories within neuroscience associated with (a) large-scale neural networks, (b) computational modeling of neural function, and (c) global workspace approaches to understanding conscious awareness. In the following section, I will then outline how these domain-general theories have been integrated and applied to emotion within the three-process model, and how that model organizes and accounts for the empirical findings discussed in the first section. Next, I describe how this model can provide a useful perspective on the processes that contribute to clinical disorders and their treatment. Finally, I will describe some implications that the three-process model may have for revising, extending, and clarifying the IMM (Lane, Ryan et al., 2015), around which the present volume is organized.

Categories of Implicit Emotion and Associated Empirical Findings

Unconsciously Caused Emotion

One category of implicit emotion—*unconsciously caused emotion*—pertains to instances in which individuals display objectively measurable (i.e., physiological, neural, behavioral) reactions to affective stimuli and may also self-report experienced changes in their own affective/emotional state; yet, they simultaneously report no experience or awareness of the stimuli that triggered those changes in their internal state. For example, an individual displaying this phenomenon might report: "I feel anxious, and my heart is racing, but I don't know why." Thus,

they are not aware of what is causing their emotional response, but they may be aware of that response after it is elicited.

There are many previous empirical demonstrations of unconsciously caused emotion (reviewed in Kihlstrom, Mulvaney, Tobias, & Tobis, 2000; R. Smith & Lane, 2016). For example, a range of studies have illustrated that, even when affective visual stimuli are presented too quickly to be perceived consciously, they nonetheless result in measurable changes in peripheral physiology and self-reported changes in affective state (e.g., Li, Zinbarg, Boehm, & Paller, 2008; Monahan, Murphy, & Zajonc, 2000). Studies in patients with blindsight—who lack subjective visual experience due to cortical damage, but nonetheless have some preserved visual pathways—have also confirmed that non-experienced affective visual stimuli lead to similar physiological and self-reported emotional effects (i.e., mediated by those preserved visual pathways; Celeghin, de Gelder, & Tamietto, 2015; Tamietto & de Gelder, 2010).

Unconsciously Represented Emotion

A second category of implicit emotion—*unconsciously represented emotion*—pertains to instances in which individuals do not report experienced changes in their affective/emotional state; yet, they display valence-specific or emotion category-specific priming effects and may also display objectively measurable (i.e., physiological, neural, behavioral) reactions to affective stimuli. For example, an individual displaying this phenomenon might (honestly) report: "I don't feel angry about what happened" in response to a normatively anger-inducing event; yet, they might nonetheless display anger-specific semantic priming effects, and they might also display increased aggression or other behavioral/physiological patterns consistent with anger.

There are currently only a limited number of empirical findings relating to unconsciously represented emotion. With respect to the unconscious representation of emotion concept categories, one study has demonstrated that subliminal presentation of guilt-related, but not sadness-related, emotion adjectives leads to increased helping behavior and reduced indulgence behavior (i.e., behaviors consistent with guilt), despite no self-reported changes in emotion/mood (Zemack-Rugar, Bettman, & Fitzsimons, 2007). Such findings suggest that the concept of guilt was represented unconsciously. Other studies have also shown that the subliminal presentation of valenced images (e.g., images of smiles vs. frowns) can promote valence-specific behavioral changes, but in the absence of any self-reported changes in emotion/mood (Winkielman & Berridge, 2004;

Winkielman, Berridge, & Wilbarger, 2005; Winkielman, Zajonc, & Schwarz, 1997). Thus, in addition to specific emotion concepts, internal representations of valence (e.g., pleasant/unpleasant, positive/negative, etc.) also appear capable of unconscious activation.

Unconscious Affective Learning

A third category of implicit emotion—*unconscious affective learning*—pertains to instances in which individuals come to display objectively measurable (i.e., physiological, neural, behavioral) reactions and may self-report affective/emotional state changes, in response to (previously) neutral stimuli, due to a statistical associations between those stimuli and other affective stimuli during past experience; yet, such individuals do not report awareness of the statistical learning processes that led to the acquired associations. For example, an individual displaying this phenomenon might report: "Every time I walk into this room I feel really uncomfortable, but I don't know why." Thus, they are aware of the eliciting stimulus (the room) and the felt response (discomfort), but they are not aware of the associative processes that have linked them together (i.e., they don't understand the connection between them).

There are many previous empirical demonstrations of unconscious affective learning (for more in-depth discussion, see Panksepp, Lane, Solms, & Smith, 2017). These are part of a larger literature on implicit learning and its neural basis (reviewed in Reber, 2013), demonstrating that (a) behavioral performance on affective learning tasks can improve, in the absence of reported awareness, following repetition and positive/negative feedback and that (b) the underlying learning processes occur within neural systems associated with motor control and habit-learning (e.g., basal ganglia) and do not depend on neural systems associated with consciousness or declarative memory. In animals, it has also been shown that both conditioned fear and conditioned taste aversion can be acquired during general anesthesia—when conscious perception of the stimuli and consciously feeling the unpleasant responses (i.e., on which the affective learning is based) would be impossible (Bermudez-Rattoni, Forthman, Sanchez, Perez, & Garcia, 1988; Burešová & Bureš, 1977; Millner & Palfai, 1975; Pang, Turndorf, & Quartermain, 1996; Roll & Smith, 1972; Rozin & Ree, 1972). Therefore, after such unconscious learning processes occur, an organism could consciously perceive a stimulus and consciously perceive an associated affective response; yet, they would have no understanding of (e.g., no accessible declarative memory accounting for) why that stimulus is leading to that affective response (for other related examples in humans, see Kihlstrom et al., 2000).

Domain-General Perspectives on Neuro-Cognitive Function

To account for the previously discussed three categories of implicit emotion, in the present section I will first outline three domain-general perspectives on the relationship between cognitive and neural functioning. In the following section, I will then illustrate how these perspectives are integrated within the three-process model and how that model can organize and provide additional insights regarding implicit emotional phenomena.

The Large-Scale Network Perspective

Based on both resting state (e.g., Yeo et al., 2011) and task-based (e.g., S. Smith et al., 2009) neuroimaging data, it has become clear that the brain's functional architecture can be usefully characterized by segregation into several large-scale networks. Such networks are comprised of spatially discontiguous, but anatomically connected (i.e., via white matter pathways; Hermundstad et al., 2013), regions that span all areas of association cortex as well as connected subcortical regions. This has led to the proposal that distinct domain-general functions can be assigned to each such network (Barrett & Satpute, 2013), and that task-specific interactions between hub regions within and between these networks may provide a useful account of more complex psychological and behavioral phenomena (Anderson, 2014).

Based on one leading proposal (Barrett & Satpute, 2013), the following networks can be assigned the following domain-general functions: (a) the salience network (SN), consisting of specific anterior insula and anterior cingulate regions (among others), plays an important role in representing homeostatic and metabolic information—based on afferent input from the body—and using that information to guide attention and behavior; (b) the default mode network (DMN), consisting of medial prefrontal, posterior cingulate, medial temporal, and lateral temporal regions (among others), plays an important role in conceptualizing the meaning of sensory input based on prior experience (also see Binder et al., 1999; Binder, Desai, Graves, & Conant, 2009); (c) the executive control network (ECN), consisting of dorsolateral prefrontal and inferior parietal regions (among others), amplifies and suppresses the strength of neural representations based on current goals; (d) the limbic network (LN), consisting of orbitofrontal cortex regions, ventral striatum, amygdala, and periaqueductal gray (among others), plays a primary role in visceromotor representation and regulation; and (e) the sensorimotor network (SMN), consisting of somatosensory cortex and

motor cortex (among others), plays a primary role in somatosensation, proprioception, and skeletomotor control. While these five networks will play a primary role in the following discussion, other relevant networks include the dorsal attention network (DAN; linked to visuospatial attention) and the visual network (VN; linked to visual perception).

The Computational Perspective

At a different level of description, neural processes have also been usefully modeled in computational terms. Computational neuroscience is a large and diverse area of research, and here I will focus only on two subareas: predictive processing (PP) models and reinforcement learning (RL) models.

Predictive Processing Models

PP models describe the brain as an organ that functions to implement tractable approximations to hierarchical Bayesian inference, via a prediction-error minimization process (Friston, 2010; Friston, Stephan, Montague, & Dolan, 2014). To a first approximation, this perspective suggests that the brain's architecture can be envisioned as implementing a multilevel generative internal model, which attempts to predict each wave of sensory input before it arrives. More specifically, each level in the model attempts to predict the pattern of activity at the level below (while also modulated by other patterns of activity at the same level; i.e., laterally), with the lowest level representing sensory input itself and all other levels representing probability distributions over possible interpretations of that input. For example, if the concept "baseball" was most activated at a higher level, this level might issue downward signals predicting perceptual representations of "small," "white," and "round" at lower levels. When the model's predictions (prior expectations or "priors") are incorrect, the resulting prediction-error signals are propagated both laterally and to the level above, and are used to revise the internal model so as to find a new set of representations (i.e., probability distributions over interpretations) that minimize these error signals. For example, if visual input led to perceptual representations of "large," "orange," and "round" at lower levels, these levels might convey error signals upward that would promote activation of the concept "basketball" instead (i.e., because these lower-level representations would be predicted by the presence of a basketball but not that of a baseball).

As these error signals can arise from different sensory modalities, as well as both within and between many different hierarchical levels, it is also necessary that they be dynamically weighted with respect to their estimated reliability (or "precision") in a given context. This allows the brain to selectively minimize

error (and therefore learn most from) signals that are most reliable/informative in a given context—a function often identified with selective attention (Feldman & Friston, 2010). For example, visual signals may often be more reliable, and therefore have higher estimated precision, during the day than at night. This type of context-specific weighting of synaptic connections within/between levels can also allow for distinct patterns of effective connectivity between brain regions/networks in different situations/tasks (Clark, 2015).

In the brain, it is proposed that prediction-error signals are represented by the activity of superficial (layer 2/3) pyramidal cells in each level, whereas perceptual inferences that minimize prediction error (i.e., the brain's best guesses about the causes of sensory input) are represented in the activity of deep (layer 5/6) pyramidal cells in each level; in contrast, stored model parameters (such as priors and precision estimates) are represented by synaptic weights between these neurons and are changed more slowly via Hebbian learning mechanisms (Bastos et al., 2012; Bogacz, 2017). Via these nuanced predictive dynamics, the resulting process allows the brain to simultaneously infer the probability of many different interpretations/descriptions of sensory input. At low hierarchical levels (i.e., nearer to the sensory periphery), these descriptions will include information about perceptual properties, such as the presence and location of edges, colors, shapes, phonemes, itches, pains, heartbeats, and many others. Such properties tend to involve statistical regularities in unimodal sensory input over short spatiotemporal scales. At higher hierarchical levels, the brain's internal descriptions of events will include multimodal conceptual properties that represent statistical regularities over longer spatiotemporal scales, such as the identification of categorical phenomena like chairs, friends, word/sentence meanings, emotions, goals, intentions, and many others. This inferred, probabilistic, multilevel description of one's situation can then be used by the brain to make decisions about how to act—with the most probable interpretation across levels being most closely associated with conscious perception and decision-making (Hohwy, 2014).

Reinforcement Learning Models

RL models describe the brain as an organ that learns to assign value to states of the world (e.g., being at a restaurant) and to actions within those states (e.g., ordering a salad vs. a cheeseburger) and that then uses those learned values to make decisions (Daw, Niv, & Dayan, 2005; Wilson, Takahashi, Schoenbaum, & Niv, 2014). It is often simply assumed in such models that the brain correctly perceives/interprets the state of the world it is in (i.e., the very complicated and difficult process that PP models attempt to describe). RL models then describe two broad classes of algorithms through which the brain appears to learn the values of those states and of the actions that might

be taken within them (and through which actions are subsequently selected in decision-making).

The first class of algorithms is model-based (MB). MB algorithms learn (and subsequently store) an internal model of all the different possible states of the world, their values, and the probability of transitioning from one state to the next (i.e., which may or may not be conditional on choosing certain actions; e.g., transitioning from the state of "being hungry at a restaurant" to the state of "tasting a cheeseburger" is conditional on the action of "ordering a cheeseburger," whereas transitioning from the state of "tasting a cheeseburger" to that of "feeling happy" may not depend on any action). Then, during decision-making, the long-term outcomes of different actions can be simulated using this internal model, and the brain can use this information to choose the action with the most valuable predicted long-term outcome (e.g., "If I order the cheeseburger I'll feel good now but bad in a couple hours, whereas if I order the salad then I'll feel much better later"). The major benefits of MB algorithms are that they learn quickly and that they are flexible. The major drawbacks of MB algorithms are that they are computationally very expensive and that their use becomes intractable in most real-world situations (i.e., where there are simply too many relevant possible futures to simulate).

The second class of algorithms is model-free (MF). MF algorithms simply store a single averaged value for each state and for each action that can be taken in that state (i.e., each "state-action pair"); these values are then updated slowly through experience via a reward prediction-error signal that represents the mismatch between a state or action's currently stored value, and the observed value of the state it subsequently leads to on a given occasion (e.g., "being at a restaurant" and "ordering a cheeseburger" would both end up with high stored values if they were repeatedly and directly followed by "tasting a really good cheeseburger"). Through a process called temporal difference learning, and with sufficient experience in a stable environment, each state or action's value can come to correctly approximate the (discounted) long-term reward expected when occupying that state or taking that action. Unlike with MB algorithms, however, no internal model is learned and no forward-looking simulations are possible. Therefore, a purely MF agent would simply find himself or herself with preferences for certain states and actions (due to the average long-term reward that has followed those state/actions in past experience), but that agent would not "know" anything about why he or she has those preferences (e.g., when at a restaurant, a purely MF agent would simply feel a strong desire to order a cheeseburger, but they wouldn't know why).

Experimental work suggests that something similar to both MB and MF processes co-exist in the brain and that they compete for control of behavior (Daw et al., 2005). More recently, PP models have also been extended to include a

process called "active inference" that can account for the behavioral results linked to both MB and MF processes within a single hierarchical system that controls bodily/behavioral processes via induction of prediction-error signals, and associated closed-loop control processes, within skeletomotor and visceromotor reflex arcs (Friston et al., 2016; Pezzulo, Rigoli, & Friston, 2015); this new perspective is useful in that it is able to anchor subjective value to homeostatic, metabolic, and other related variables necessary for the survival and reproduction of biological organisms (including highly social organisms like humans). It also has advantages over some RL models in that it provides a principled means of motivating agents to value exploratory behavior (i.e., promoting learning about their environments), instead of primarily motivating reward-seeking (and pain-avoidant) behavior (i.e., whereas RL models often promote directed exploratory behavior in less natural ways; e.g., adding an "exploration bonus," see Sutton & Barto, 1998). For present purposes, however, I will focus on PP models for describing perception and conceptualization processes, and I will focus on the simpler RL models for describing behavioral control processes.

The Global Workspace Model and Conscious Access to Representational Content

It is widely recognized that regions of the brain locally represent large bodies of information in parallel (e.g., probabilistic information with PP models) and that the vast majority of such information does not contribute to conscious experience in a given moment. One leading model aiming to describe the neural processes that determine what representations do (and do not) contribute to conscious experience from moment to moment is called the global workspace model (reviewed in Dehaene, 2014; Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006; Dehaene, Charles, King, & Marti, 2014). Briefly, this model assumes that, via lateral inhibition, many different representations (activated in parallel by sensory input) locally compete for access to the brain's "global neuronal workspace"—a set of central, highly connected hub regions within large-scale neural networks (across frontal, parietal, and temporal cortex; often referred to as "rich-club hubs"; van den Heuvel & Sporns, 2013) that facilitate global transmission/availability of represented information. The outcome of the aforementioned competition is determined by a combination of bottom–up stimulus strength and top–down influence from cognitive control processes (e.g., involving the ECN and DAN in large-scale network models). When a representation "wins" this competition, this corresponds to the initiation of a nonlinear amplification process that engages strong reciprocal interactions between that representation and the global workspace. This allows the content of that representation to have

a widespread and synchronous influence on large-scale brain functioning (i.e., widespread and efficient internal model updating) and further allows it to be maintained and manipulated within working memory over long-time scales (i.e., and therefore contribute to the deliberative, forward-looking decision-making processes associated with MB algorithms). It is this global influence of a "winning" representation's content that correlates with conscious experience and self-reported awareness (also see R. Smith, 2016, 2017). Thus, while large amounts of information are represented locally, only a subset of this information is allowed to become conscious and therefore have a more global influence on serial, multi-step cognitive processes associated with goal-directed decision-making.

The Three-Process Model of Emotion Episodes

The three-process model (TPM; see Figure 3.1; Lane, Weihs, Herring, Hishaw, & Smith, 2015; Panksepp et al., 2017; R. Smith, Killgore, et al., 2017; R. Smith & Lane, 2015, 2016; R. Smith, Thayer, Khalsa, & Lane, 2017) seeks to integrate each of the three perspectives described in the previous section to provide a satisfactory theoretical account of the various categories of implicit emotion as previously discussed. I will next discuss each process in this model.

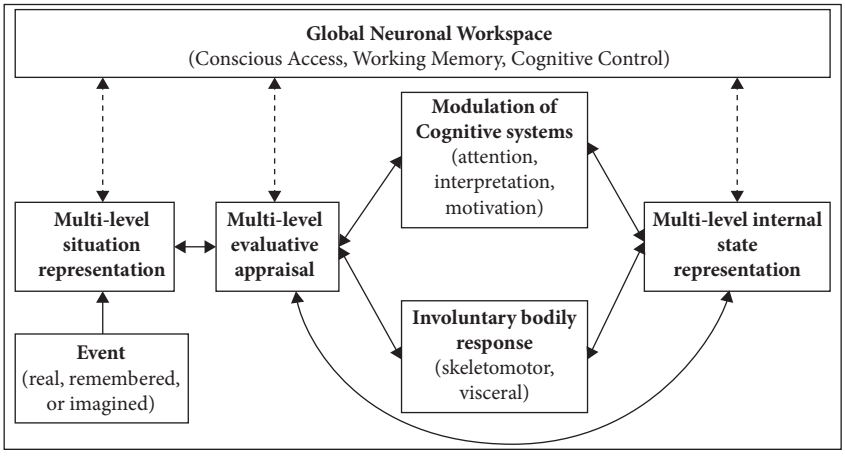

Figure 3.1 Affective response generation (ARG) processes: interactions between multi-level evaluative appraisals, modulation of cognitive systems, and involuntary bodily responses. Affective response representation (ARR) processes: multi-level internal state representation. Conscious access (CA) processes: interactions between the global neuronal workspace and other model elements (indicated by dotted arrows).

According to the TPM, an emotion episode is initiated when an event (whether real, remembered, or imagined) is represented within the brain, in the hierarchical and probabilistic fashion suggested by PP models. This corresponds to the "multilevel situation representation" box in Figure 3.1. Expanded out, this box would contain representations of the low-level perceptual features of the event (represented within cortical sensory systems) as well as higher-level (and longer timescale) conceptual descriptions of the event (i.e., associated with the conceptualization processes linked to the DMN and its interactions with long-term episodic and semantic memory systems). If/when represented aspects of this description are selected for global broadcasting, the content of those representations would be experienced as a consciously perceived, remembered, or imagined event. Such descriptions could also be given a verbal gloss such as, for example, "My co-workers didn't invite me to lunch because they do not like me."

Affective Response Generation Processes

Once such a multilevel representation is in place, the TPM suggests that this representation is then probabilistically evaluated for its significance to one's own needs, goals, and values across a large number of appraisal dimensions (e.g., as supported by a large body of work on appraisal theories of emotion (e.g., see Scherer, 2009; Siemer, Mauss, & Gross, 2007). This includes, for example, evaluations of a represented event's expectedness/novelty, its relevance/importance, its congruence/incongruence with one's goals, its consistency with one's norms/values, its controllability, and whether responsibility for the event belongs to the self or to others (corresponding to the "multilevel evaluative appraisal" box in Figure 3.1). Depending on the outcome of this high-dimensional evaluation, appropriate changes are initiated in the state of the body and to the state of other cognitive systems; these changes reflect the predicted metabolic, cognitive, and behavioral demands of the situation *as evaluated*. For example, if the previous situation "My co-workers didn't invite me to lunch because they do not like me" was evaluated as unexpected, goal-incongruent, controllable, and high in other-responsibility, then a high arousal, negatively valenced bodily response might be initiated to prepare for aggressive actions aiming to change the situation (e.g., going to yell at your co-workers); such a reaction would also be accompanied by associated biases in attention, memory, interpretation, and action selection (e.g., see Chapter14 in Shiota & Kalat, 2012). In contrast, if the same situation were appraised/evaluated differently, then a different response would be generated.

It is important to highlight that, in addition to the previously described cognitive evaluation, lower-level predictive/associative processes can also link event

representations to such bodily/cognitive responses more directly. For example, if a given representation (e.g., of being in a particular room) has repeatedly been paired with an aversive outcome (e.g., pain), then prediction-error minimization processes would lead the "room" representation to predict the "pain" representation, which can then elicit pain-related cognitive/bodily responses directly (i.e., a classically conditioned response). Thus, associative learning processes (e.g., such as classical and operant conditioning processes) can contribute to bodily/cognitive reactions, and urges to act in particular ways (e.g., via MF algorithms), in addition to (and sometimes in competition with) the influence of higher-level cognitive evaluation.

Combining large-scale network models and PP models, the TPM envisions these affective response generation (ARG) processes to be implemented in the brain as follows. First, the sensory input associated with an event, internal (e.g., a decrease in glucose) or external (e.g., not getting invited for lunch), would meet current model predictions and generate an array of prediction-error signals that, when weighted by attentional modulation (i.e., based on precision-estimates and implemented via modulation of postsynaptic gain), would lead to an updated internal description of the event that approximates optimal probabilistic (Bayesian) inference. The low-level features of the description would rely on cortical sensory systems, whereas the higher-level conceptual features would rely on DMN-mediated conceptualization processes (See Figure 3.2). The evaluation of cognitively complex appraisal dimensions would also rely on these DMN processes, whereas more direct associative links (e.g., conditioned responses) would involve interactions between cortical sensory systems and the LN-mediated visceromotor control processes. Ultimately, the LN would initiate a visceromotor response, and interact with subcortical neuromodulatory nuclei (e.g., cholinergic, noradrenergic, dopaminergic, and serotonergic nuclei; see Chamberlain & Robbins, 2013; Cools, Nakamura, & Daw, 2011) to alter cognitive/attentional/motivational biases, in response to predictive signals received from both cortical sensory systems and from the DMN. The SMN and subcortical nuclei subserving skeletomotor control (e.g., basal ganglia, facial motor nucleus, etc.) are also plausibly involved in the initiation of associated facial expression and body posture changes, and the SN is also implicated in generating autonomic responses. (For a review of evidence supporting the envisioned neural basis of these processes and those described further in the following discussion, see R. Smith, Killgore, et al., 2017).

Affective Response Representation Processes

After the previously discussed processes have generated this bodily/cognitive reaction, the TPM suggests that a probabilistic, constructive process (also see Barrett, 2017) is required to subsequently represent that reaction and to infer its

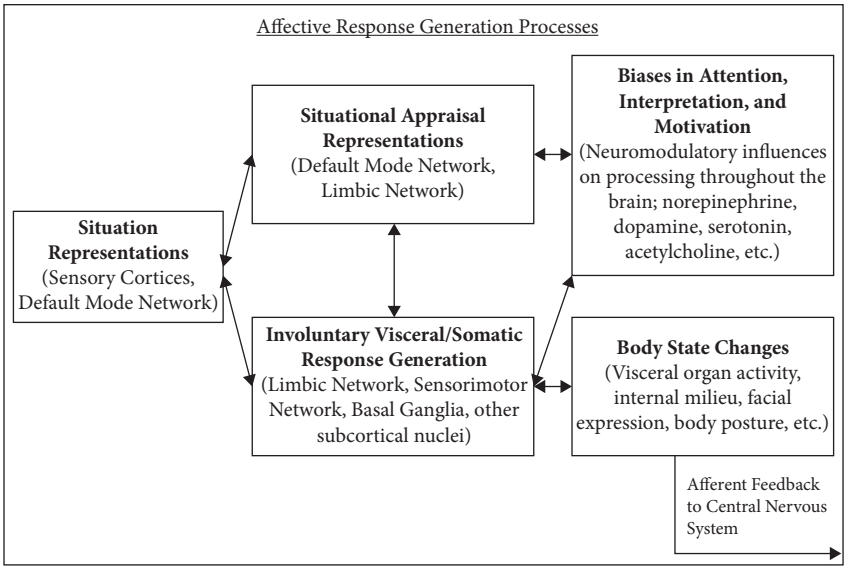

Figure 3.2 Schematic illustration of the affective response generation (ARG) processes (and their neural basis) described in the text. Bi-directional arrows indicate the exchange of prediction and prediction error signals, leading to internally represented probability distributions over possible states (e.g., possible situation representations, possible visceral/somatic states, etc.). Biases in attention, interpretation, and motivation also involve modulation of postsynaptic gain, based on estimates of precision/reliability.

emotional meaning (i.e., corresponding to the "multilevel internal state representation" box in Figure 3.1). This can be expanded out to include both perceptual and conceptual levels of representation.

At the perceptual level, this will involve updating representations—via the same previously described PP model dynamics—of how the state of the body has changed (i.e., as filtered through the cognitive biases in attention/interpretation that were simultaneously initiated). If/when such updated body state representations become selected for global broadcasting, they would also contribute to experienced changes in heart rate, respiration, muscle tension, facial expression, body posture, subjective energy level, and a range of other variables. Within large-scale network models, this is envisioned to involve afferent prediction-error signals, arising from the body, that are used to update internal estimates of such variables within the SMN (i.e., skeletomotor variables) and the SN/LN (i.e., visceral variables).

At the conceptual level, this will involve updating representations of the learned concepts that best describe the current felt state of the individual. In the

case of affective responses, the most relevant class of concepts is that associated with emotion categories (e.g., anger, fear, happiness, sadness, etc.)—which make joint predictions about both internal and external variables (for specific predictive processing models for emotion concepts, see Smith, Lane, et al., 2019; Smith, Parr, & Friston, 2019). For example, an individual's learned concept of sadness might predict internal feelings of low energy and the desire to be alone, as well as external descriptions of situations involving the loss of something valued, low controllability, and high personal importance. In contrast, an individual's learned concept of anger might instead predict internal feelings of high arousal, the desire to respond aggressively, and situations involving goal frustration and high levels of other-responsibility.

The TPM suggests that DMN-mediated conceptualization processes (modeled as prediction-error minimization processes) can be envisioned as converging on a (probabilistic) representation of the emotion concept or concepts that best predict (and therefore minimize prediction-error with respect to) the pattern of representations of one's internal and external situation implemented elsewhere in the brain. This will include both lateral interactions within the DMN (i.e., between emotion concepts and situation conceptualizations) and hierarchical interactions with other neural networks, including the SMN/SN/LN (i.e., representing bodily variables) and exteroceptive cortical sensory systems (e.g., visual/auditory cortex; i.e., representing external situational variables). Thus, for example, the DMN might converge on a representation that the concept of fear has the highest probability of accounting for a particular situation, which involves perceptions of high heart rate, strong desires to run away, and the presence of a dangerous animal. If/when this representation of the concept of fear was selected for global broadcasting, it would be experienced as the conscious recognition that one felt afraid.

Conscious Access Processes

The previous discussion highlights a large number of representations relevant to the course of an emotion episode. These representations include perceptual and conceptual representations of an event (whether real, remembered, or imagined), representations of evaluative appraisals of that event, representations of bodily responses initiated in reaction to those appraisals, and conceptual representations of one's current state (e.g., emotion concepts like sadness or fear). The TPM suggests that all of these representations compete for access to the global neuronal workspace via lateral inhibition and that only the content of "winning" representations will be consciously experienced from moment to moment. Such a competition will be biased by bottom–up stimulus strength (e.g.,

intense increases in heart rate will be more likely to enter conscious experience than will mild increases); it will also be influenced by top–down filtering processes (e.g., one will be more likely to consciously recognize their own emotions if they have the goal of attending to what emotion they are feeling—mediated by the influence of the ECN). When a representation wins the competition for global broadcasting, this is envisioned as an increase in bidirectional effective connectivity between the set of neurons implementing that representation and the distributed hub regions of the global neuronal workspace. This increase in bidirectional effective connectivity causes that representation to be amplified/maintained within an active state and allows it to contribute to the multistep, goal-directed planning processes associated with working memory and MB decision-making.

However, even if a representation is not selected for global broadcasting, it will nonetheless be capable of influencing behavior via direct predictive/associative links. For example, certain situation representations, appraisals, body state representations, or emotion concept representations might directly prime certain actions due to past reinforcement history (i.e., MF learning), even if the content of such representations does not enter consciousness. Spreading semantic activation could also lead concepts to influence cognition in a semantic priming task (e.g., the unconscious activation of the concept "fear" might speed up reaction times for identifying negatively vs. positively valenced words).

Overall, the TPM allows for many combinations of experienced/unexperienced aspects of an emotion episode (see Figure 3.3). For example, if situation/appraisal representations *were not* selected for global broadcasting on a given occasion, but emotion concept representations *were* selected for global broadcasting, this would correspond to the category of *unconsciously caused emotions* described in the first section on categories of implicit emotion and associated empirical findings. In contrast, if situation/appraisal representations *were* selected for global broadcasting on a given occasion, but emotion concept representations *were not* thus selected, this would correspond to the category of *unconsciously represented emotions* described in the first section. Finally, if *neither* situation/appraisal representations *nor* bodily/emotion-related representations *were* selected for global broadcasting on a given occasion, this would still engage the implicit predictive/associative learning process associated with both PP and RL models, and therefore promote future affective reactions that are poorly understood (i.e., the *unconscious affective learning* processes in the first section).

These are just three examples of a much broader range of possibilities, in which any combination of situation representations, appraisal representations, interoceptive/somatic representations, and emotion concept representations might or might not be experienced in a given situation—depending on the

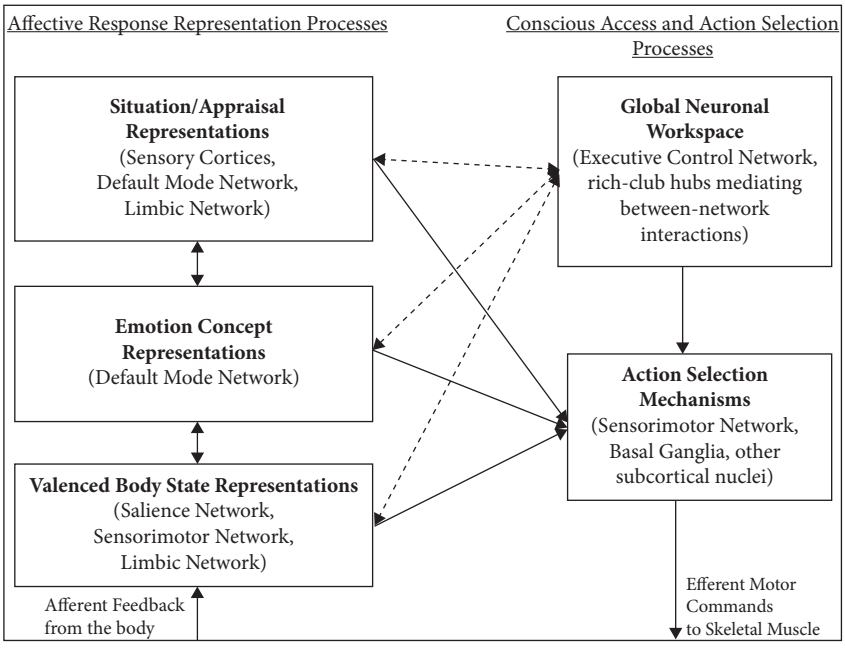

Figure 3.3 Schematic illustration of the affective response representation (ARR) processes and conscious access (CA; and their neural basis) processes described in the text. Bi-directional arrows indicate the exchange of prediction and prediction error signals, leading to internally represented probability distributions over possible states (e.g., possible emotion concept representations, etc.). Dashed arrows indicate information flow that depends on a representation being selected for global broadcasting, leading to conscious experience and flexible use in goal-directed decision-making processes. Solid arrows between representations (left) and action selection mechanisms (bottom right) indicate the implicit influence such representations can have on action selection in the absence of conscious access (e.g., the influence of model-free state-action pair values).

combination of stimulus intensity and goal-directed attention (Dehaene et al., 2006). In previous work, for example, I (in collaboration with other colleagues) have described how an individual might consciously perceive their own affective/bodily response (e.g., lethargy) to a meaningful event (e.g., the loss of a loved one) and yet fail to recognize the emotional meaning of that bodily response (e.g., fail to represent, or gain conscious access to, the fact that "sadness" best describes this felt change in state; see Lane, Weihs, et al., 2015; R. Smith & Lane, 2015, 2016). We have referred to such phenomena as involving an "affective agnosia," comparable to similar visual recognition deficits in associative visual agnosias.

Relevance to Clinical Practice

As illustrated in the previous section, the TPM can account for the various categories of implicit emotion described in the first section. It also highlights what conscious access allows, including the ability to hold information in working memory and use it to guide goal-directed cognition and action selection. Beyond these insights, the TPM also makes contact with various other empirical/theoretical phenomena of clinical interest. In this section, I will discuss a few major examples of these points of contact and their implications.

Individual Differences in Trait Emotional Awareness

One area where the TPM has resources capable of providing important potential insights pertains to individual differences in tEA (Kashdan, Barrett, & McKnight, 2015; Lane & Schwartz, 1987). Individuals with low tEA describe their emotions in physical (e.g., "I feel sick to my stomach") and/or coarse-grained (e.g., "I feel bad") terms and have difficulty differentiating their own emotions from those of others, whereas those with high tEA describe their emotions in highly granular psychological terms (e.g., "I feel a mix of fear and jealousy") and easily differentiate their feelings from those of others. Low levels of tEA have been associated with a range of psychiatric conditions (reviewed in Lane, Weihs, et al., 2015), and higher levels of tEA have also been shown to facilitate better outcomes in panic disorder across both cognitive-behavioral therapy (CBT) and manualized psychodynamic psychotherapy (Beutel et al., 2013; for other positive benefits of emotion differentiation ability, see Kashdan et al., 2015). However, the neurocognitive basis for differences in tEA remains poorly understood.

In a recent article, I (in collaboration with other colleagues) proposed that the TPM allows for three different (and potentially complimentary) explanations for individual differences in tEA (R. Smith, Killgore, et al., 2017). First, individuals who make use of more appraisal combinations to describe the events in their lives would also be expected to describe a wider range of emotional experiences (i.e., because more unique appraisal combinations would be expected to lead to the generation of more unique affective responses). Thus, stable differences in these ARG processes could explain measured differences in tEA. Second, individuals who have learned a wider range of emotion concepts (and/or richer scripts/schemas associated with those concepts) would also be expected to describe a wider range of emotional experiences (e.g., a person could not describe feelings of jealousy if they have not learned the concept of jealousy). Thus, stable differences in this aspect of affective response representation (ARR) processes could explain measured differences in tEA as well. Finally, individuals who have

learned to value information about emotions, and who have acquired goals for which information about emotions is highly relevant, would also be expected to become aware of information about emotion more often. This is because such individuals would attend to their own emotions more frequently (i.e., due to their higher expected relevance/value), and this would lead emotion-related representations to enter the global neuronal workspace on a broader range of occasions. In contrast, other individuals may instead have had experiences that reinforce avoidant patterns of attention (as described in more detail elsewhere; R. Smith & Lane, 2016). If attention to emotion repeatedly leads to increased discomfort for an individual (e.g., if they believe their own emotion is unacceptable for some reason), then avoidant attention would also reduce this discomfort and could become habitual over time (i.e., MF algorithms would learn higher values for the "action" of attending *away* from one's own emotions)—leading such a person to report awareness of fewer emotions. Thus, stable differences in CA processes could also explain measured differences in tEA.

At present, these different potential explanations remain to be tested. However, it is worth highlighting that they each entail a specific cognitive process that could be targeted by interventions aiming to improve tEA in clinical populations. For example, if ARG process differences account for variance in tEA levels, this would suggest that future efforts to improve tEA could focus on training individuals to make more nuanced situational appraisals (e.g., to reduce "black and white thinking" about the situations they are in). In contrast, if ARR process differences account for variance in tEA levels, it would suggest tEA training efforts might also focus on increasing the richness of emotion concept knowledge (e.g., to reduce the use of only coarse-grained, non-specific concepts like "bad" and "good" to understand emotional experience). Finally, if CA process differences account for variance in tEA levels, this would suggest that tEA training efforts could further focus on increasing the value individuals place on, and the amount of attention they therefore allocate to, emotional information (e.g., if individuals believe information about emotions is valuable/useful to their personal needs and goals, they will likely spend more time attending to and reflecting on this information). However, it should be recognized that the correct way of implementing such possible interventions in clinical populations (i.e., depending on which cognitive process differences are empirically linked to tEA in future research) also remains an important open question. For example, as avoidance processes (e.g., the previously mentioned MF avoidant attention processes) could plausibly prevent some individuals in clinical populations from being receptive to learning new information (and fully participating in such training efforts), the implementation of these interventions may need to be designed in a manner that could overcome such defensive processes in such individuals.

Conceptualizing Emotional Pathology

The TPM also highlights multiple ways in which a lack of awareness may contribute to chronic emotional pathology. For example, lack of conscious access to relevant situation/appraisal representations, as in unconsciously caused emotions, can interfere with an individual's ability to identify and resolve the factors maintaining negative affect. Further, lack of conscious access to emotion concept representations, as in unconsciously represented emotions, can further prevent an individual from understanding why they are experiencing unpleasant bodily sensations—and this poor understanding can even lead to further sources of anxiety/arousal (e.g., mistaking panic symptoms for signs of a heart attack). As high levels of arousal interfere with goal-directed cognition (Teigen, 1994), this can also prevent subsequent reflection (e.g., model-based reasoning processes) that might otherwise serve to regulate it. Previous work on the role of low tEA (and affective agnosia) within psychiatric conditions also suggests that lack of conscious access to the previously mentioned types of information may contribute to multiple disorders (for reviews, see Lane, Weihs, et al., 2015; R. Smith & Lane, 2016).

The TPM also offers a framework within which it is possible to theorize about the origins of other types of clinical symptoms. For example, the computational level of description within the TPM strongly emphasizes the influence of prior experience and prior expectations on each of the inferential processes involved in emotion episodes. For example, to infer the most likely description of a new pattern of sensory input (e.g., the meaning of a facial expression, the meaning of a felt bodily sensation, etc.), the brain must use the current predictions of its internal model (i.e., its priors) as a starting point and then adjust internal representations from this starting point to minimize prediction error. As such, if two people started out with different priors at the time of perceiving an event (e.g., elevated heart rate), they may end up perceiving/understanding that same event in two different ways (e.g., "I'm anxious" vs. "I'm excited"). Differences in past experience may similarly lead two people to have different estimates regarding the context-specific reliability (i.e., precision) of different types of sensory inputs—further leading such individuals to attend to (and ignore) different features of the same overall experience.

By biasing how sensory input is interpreted and responded to, this influence of prior expectations and precision estimates could play an important role in the maintenance of some types of emotional pathology. One important example of this pertains to the acquisition of overly strong (precise) priors for particular appraisal patterns, which, according to the TPM model, could bias individuals toward having recurring, contextually inappropriate affective responses. For example, if an individual had learned an abnormally strong (high-precision) and

generalized expectation for high levels of other-responsibility, they might be expected to experience overly frequent/intense anger across a range of contexts. In contrast, if an individual had learned an abnormally strong (high-precision) and generalized expectation for high levels of self-responsibility, they might instead be expected to experience more frequent/intense guilt across a range of contexts (for further discussion of such "appraisal biases," see Scherer, 2009). More generally, overly strong priors for any type of abstract outcome would be expected to bias attention toward congruent information and promote congruent interpretations of ambiguous stimuli. As explored elsewhere (R. Smith, Akozei, Killgore, & Lane, 2017), this type of process is likely to play an important role in explaining how particular schemas in depression maintain negativity biases and continue to promote normatively inappropriate negative reactions to ambiguous situations.

At the psychological level of description, the TPM also clarifies how unconscious/implicit emotion can (and cannot) plausibly be understood to promote/maintain emotional pathology. One key point to emphasize here is that, in the TPM, affective responses are generated, maintained, adjusted, and regulated *from moment to moment* as the brain (a) continues to receive new interoceptive/exteroceptive sensory input and (b) continues to update its internal descriptions/appraisals of the meaning of that input. The competition between activated representations for conscious access is also a dynamic process, such that the contents of consciousness are similarly updated/maintained on a moment-to-moment basis. As such, it is not plausible to think of the unconscious as a "place"; nor is it plausible to think of emotions as "things" that are "sitting in the unconscious" waiting to be discovered. Instead, statements about an individual with an "unconscious emotion" can only be understood to refer to facts about that individual's *currently* generated affective responses, and what emotion concepts may *currently* be primed by those responses, in the absence of reportable awareness by that individual.

This conception of the unconscious—not as a place but instead as a way of describing the moment-to-moment global accessibility of represented information—suggests that some clinical phenomena that *appear to* involve unconscious emotion may instead involve other processes, such as the generation of new affective responses and/or new conceptualizations of those responses. For example, consider a hypothetical client who suffered abuse by a parent in early childhood. Such an individual may report awareness of fear as a child, but no awareness of parent-directed anger, and yet this individual may come to report awareness of such anger during psychotherapy as an adult. In such cases, one might suggest that this anger was actually "in the unconscious" since that individual's childhood, and that letting it "become conscious" in therapy contributed to that individual's growth/progress. However, the TPM requires a different way of understanding such cases. One reason for this is that, while abuse in

childhood surely led to the generation of negatively valenced, high arousal states (likely also involving avoidant/escape-related action priming), such affective responses are not plausibly maintained continuously throughout an individual's life; further, those responses were reportedly conceptualized as fear, and no other available information suggests the concept of anger was also activated unconsciously to represent the meaning of those responses *at the time*. Therefore, according to the TPM, it is more plausible that, during the course of therapy, the individual was led to view the memories of those events from a different perspective (e.g., using the social norms/expectations learned in adulthood), leading those events to be understood/appraised differently. These new appraisals could then lead to the generation of a *new* affective response in that individual (likely involving aggressive/approach-related action priming), which that individual subsequently conceptualizes and reports as anger. So what appears to be the uncovering of a pre-existing emotion in such cases may instead involve the generation of a new emotion in response to the memory of the (previously fear-inducing) event, and it is the generation of this new emotion (and the associated new appraisals of the memory) that contribute to therapeutic progress. This example therefore serves to illustrate how, from the perspective of the TPM, some clinical phenomena that are often thought of as involving unconscious emotion will likely need to be reconceptualized (for similar suggestions in previous work, see Lane & Garfield, 2005; Lane, Weihs, et al., 2015).

Conceptualizing Therapeutic Mechanisms

The TPM also provides resources for conceptualizing a range of therapeutic mechanisms. For example, when situation/appraisal representations become conscious, this allows the maintenance/manipulation of those representations within goal-directed cognition—a process that appears central to the cognitive reappraisal strategies taught within CBT (Buhle et al., 2014). Within the TPM, reappraisal can be thought of as a goal-directed (MB) process in which event interpretations are maintained/manipulated within working memory, with the aim of finding plausible alternative interpretations. When such alternative interpretations are found, ARG processes automatically evaluate/appraise them—potentially leading to the generation of a different and more adaptive affective response. This also plausibly allows such event memories to become associated with (i.e., come to predict) these new affective responses, and could change the stored valence of those memories. Note, however, that this process of altering the concept-level interpretation of one's situation may be less effective if the clinically relevant ARG processes are primarily driven by lower-level perceptual representations (e.g., where particular perceptual features, like being "black and

fuzzy," are associatively linked to a particular affective response). In such cases, treatments that instead facilitate new experiences with the relevant percepts (e.g., exposure therapies; discussed further in the following text)—and therefore allow MF processes to slowly alter the strengths of previously learned associative (i.e., Pavlovian) links between percepts and affective responses—would be expected to be more effective.

As another example, when emotion concept representations become conscious, this similarly allows the maintenance/manipulation of those representations within goal-directed cognition. This process plausibly underlies an individual's ability to reflect on their emotions in a given situation and on how they should respond. This type of process may be important in allowing individuals to (a) infer the likely causes of their emotions, (b) identify their primed (i.e., MF) action tendencies, and (c) engage MB decision processes capable of evaluating the most likely outcomes of choosing different actions. Such emotion-focused reflective processes are also an important component of multiple types of psychotherapy, as the absence of such reflection can sometimes contribute to maladaptive behavior. Consciously accessing and reflecting on emotions might also play a role in the amplification of adaptive associative learning processes in psychotherapy. That is, as conscious access to emotion concept representations (and associated percept-level body state representations) involves a top–down amplification/maintenance process, this would be expected to influence any implicit/associative learning processes that depend on the strength/duration of such activated representations. For example, emotion-focused therapy involves evoking new conscious emotional experiences in a client while they reflect on thoughts and memories of traumatic (or otherwise problematic) life events, leading to new, more adaptive affective responses when reflecting on those thoughts/memories in the future (Greenberg, 2010); it is possible that the associative learning processes that underlie this change in automatic responding would proceed much less efficiently (if at all) in the absence of these new emotional experiences (e.g., if somehow those new affective responses were activated and represented, but were never amplified/maintained within the global neuronal workspace). Therefore, in addition to facilitating more adaptive goal-directed problem-solving processes, conscious access to emotion may beneficially alter implicit learning as well (i.e., by increasing the strength/duration of activated representations, which would in turn be expected to amplify associative learning with respect to those representations).

Yet another example pertains to understanding the efficacy of acceptance-/mindfulness-based therapeutic approaches such as acceptance and commitment therapy (ACT; e.g., see Hayes & Smith, 2005). Within the TPM, such approaches can be understood to work (in part) by altering the way that represented aspects of an affective response are subsequently appraised—thus altering the

subsequent "second-order" affective responses that are generated *in response to the initial affective response* (i.e., as mediated by the U-shaped arrow in Figure 3.1 that leads from the "multilevel internal state representation" box to the "multilevel evaluative appraisal" box). For example, if an individual is currently representing their present emotional state as involving "intense anxiety," and this internal state description is then appraised as goal-incongruent and inconsistent with their own norms (e.g., they believe the state is "unacceptable," "intolerable," and/or that they "should not be feeling that way"), then this would be expected to lead to a further affective response that simply amplifies their felt anxiety. In contrast, if these therapeutic processes allow an individual to instead appraise their anxiety as goal-congruent and consistent with their own norms (e.g., they instead believe their anxiety is "acceptable," "tolerable," and/or that "it is OK to be feeling that way"), then this would *not* lead to a further affective response that would amplify their anxiety. It is therefore fairly clear how the mechanisms underlying such acceptance-based approaches can be understood within the TPM and how they would facilitate therapeutic progress.

Finally, it is well known that the computational perspective that the TPM draws from also provides a straightforward way of understanding the effectiveness of exposure-based therapies (Foa, Hembree, & Rothbaum, 2007). In such therapies, individuals remain exposed to stimuli that trigger intense negative affect for a long enough period of time that prediction-errors—arising due to the absence of the aversive events associated with those stimuli—are capable of updating an individual's internal model of the world, leading representations of those stimuli to no longer predict danger or prime avoidance responses. At the same time, it is also worth highlighting that the hierarchical aspect of processing within the Bayesian computational perspective has also been used to successfully account for puzzling phenomena in this domain—such as the fact that negative affective responses often return after being extinguished during exposure therapies. Briefly (and somewhat simplified), the aforementioned account suggests that the brain can learn to expect different associations between stimuli (i.e., predictions at a lower level) in different contexts (i.e., predictions at a higher level), and that "fear extinction" processes may typically lead the brain to infer that it is in a new context—leaving previously learned associations within the old context relatively unchanged (for a review of this explanation, and the confirmed empirical predictions supporting it, see Gershman, Norman, & Niv, 2015). This has led to the interesting (and now confirmed) prediction that gradual extinction processes will lead to revision of the original implicit "memory" (i.e., stimulus–response association) and therefore prevent the later return of affective responses, whereas abrupt extinction processes will instead lead to a new implicit memory (i.e., a new stimulus–response association, linked to a new context), and therefore allow the return of extinguished responses.

Implications for the Integrated Memory Model

The IMM, around which this volume is organized, suggests that many (possibly all) psychotherapies are effective because they lead to memory reconsolidation processes and that these revised memories involve three major components: episodic memories, semantic memories, and emotional responses. In a broad sense, the TPM is quite consistent with the IMM; in fact, recent work supports the role of prediction and prediction-error signaling in each of these memory components and within reconsolidation processes (Greve, Cooper, Kaula, Anderson, & Henson, 2017; Henson & Gagnepain, 2010; Pezzulo et al., 2015; Sevenster, Beckers, & Kindt, 2014). However, when zooming in on the details, the TPM also suggests that (a) the "emotional responses" component of the IMM requires some conceptual expansion/clarification and (b) the IMM may need to more explicitly incorporate action-value memories associated with MF algorithms. I will expand on each of these points next.

The Meaning of Emotional Responses Within the IMM

As the TPM makes clear, "emotional responses" cannot plausibly be understood as completely separate and distinct from episodic and semantic memory processes. For example, in the TPM, priors derived from episodic/semantic memory (e.g., schemas, scripts) are first used by the DMN to conceptualize and appraise the significance of an event, leading to the generation of conceptualization-/appraisal-dependent visceral/somatic responses (and linked changes to cognitive/motivational biases). Such responses are then perceived and represented as having a particular valence and signifying a particular emotion concept—and these perception/conceptualization processes also draw on priors from episodic/semantic memory. Therefore, the meaning of "emotional responses" in the IMM is unclear. One suggestion might be to replace "emotional responses" with "visceral/somatic responses" in the IMM; in this suggested revision, the activation of particular episodic memories, and particular concepts/schemas/scripts within semantic memory, would simply be associated with (i.e., predict)—and therefore initiate—particular changes in one's visceral/somatic state, and such changes would then be perceived as having valence and conceptual emotional meaning (i.e., via subsequent further interaction with episodic/semantic systems).

It is also worth highlighting that the IMM does not propose a concrete mechanism to account for when such representations of valence and emotional meaning will and will not become consciously experienced/recognized. As the TPM does propose a concrete mechanism accounting for differences in

conscious access to representational content, it may therefore help to unpack the origin of a wide range of clinical phenomena involving individuals' awareness, and lack of awareness, of different aspects of their own affective responses.

The Need for Action-Value Memories Within the IMM

Unlike the TPM, the IMM does not explicitly incorporate current work on algorithmic processes (e.g., RL processes or active inference processes) required for adaptive control of behavior. Further, while scripts/schemas within semantic memory are plausibly used to derive the priors/predictions within PP models that govern perception/conceptualization of sensory input, it is less plausible that pathology in such semantic structures can directly account for all pathological decisions/behaviors that serve to maintain psychiatric symptoms. For example, consider an individual who once had a panic attack on a train and now continually avoids trains—causing significant practical problems and distress associated with getting to work on time. I would suggest that this behavior is not plausibly accounted for only because the person has learned a semantic "script" structure about avoiding trains (or about trains leading to panic attacks). Instead, I would suggest that such script-based expectations must further interact with MF learning processes. For example, assume that each time the person approaches a train (or considers entering a train) these expectations initiate a negatively valenced, high arousal visceral state (e.g., anxiety), and then the individual chooses the action of "avoiding the train" (e.g., they decide to walk to work instead)—leading to a reduction in that negatively valenced state. Each time this happened, MF algorithms would increase the value of the "avoiding the train" action (due to negative reinforcement)—leading to that action being even more likely to be selected again the next time. The person would therefore end up having a stronger and stronger automatic, non-reflective tendency or drive to avoid the train (i.e., requiring greater and greater amounts of cognitive control to overcome). These types of MF action-value memories are not plausible elements of episodic or semantic memory within the IMM—and yet they certainly represent a type of memory that needs to be updated (e.g., via exposure and response prevention to riding trains) as a part of effective treatment.

In the TPM, MF action-value memories play a central role in accounting for the automatic/associative action tendencies that people have during affective responses—especially when represented elements of those affective responses remain outside of awareness. As the previous example illustrates, these action values are learned and they plausibly play a major role in clinically significant avoidance behaviors that prevent therapeutic progress (also

see Barlow et al., 2011). Therefore, I would suggest that the IMM should be expanded to incorporate this type of action-value learning/memory, because it is only through interactions between episodic/semantic memory processes, visceral/somatic response generation processes, and such action-value learning processes that many patterns of pathological behavior will be fully accounted for. More generally, it also appears plausible (at least to this author) that, in many cases, changes in action-value memories and changes in maladaptive priors/expectations (along with changes in learned associations between percepts, concepts, and visceral/somatic responses) could be central for effective change in psychotherapy. Thus, reconsolidation of these memories could (at least in some cases) be more relevant than, for example, reconsolidation of the content of particular episodic memories in a person's past experience. These considerations also highlight why it would be useful if the IMM were further articulated such that it was clearer how reconsolidation of the various aspects of memory (i.e., changing the content of episodic memories, learning new concepts/schemas/scripts in semantic memory, changing associations between episodic/semantic memories and visceral/somatic responses, etc.) are envisioned to account for the change process. As all types of memory within PP (and related neurocomputational) models are ultimately grounded in synaptic weights that store internal model parameters at different hierarchical levels (e.g., priors and precision estimates), presumably a fully articulated version of the IMM would clarify which internal model parameters (synaptic weights) are altered during reconsolidation (and in relation to which types of memory), as well as why/how they are altered in particular ways, as a result of effective therapeutic processes.

Conclusion

In summary, the TPM suggests multiple future directions for the IMM, and it offers many useful and clinically relevant conceptual resources. It offers a way of accounting for distinct types of implicit emotional phenomena that have been empirically observed, and it offers a way to map such phenomena onto processes at the cognitive, computational, and neural level of description. It also constrains the uses of the term "unconscious emotion" to those that can be accounted for by plausible neural mechanisms and offers a range of hypotheses regarding possible sources and maintenance factors associated with emotional pathology, as well as possible mechanisms through which therapies may affect change. While informed by a large body of previous work, all such possibilities remain to be thoroughly tested. It is this author's hope that such promising opportunities for future research on the TPM and IMM will be taken soon.

Table 3.1 Abbreviations

Abbreviation	Definition
TPM	Three-process model
ARG	Affective response generation
ARR	Affective response representation
CA	Conscious access
RL	Reinforcement learning
PP	Predictive processing
MB	Model-based
MF	Model-free
IMM	Integrated memory model
TEA	Trait emotional awareness
ECN	Executive control network
DMN	Default mode network
SN	Salience network
LN	Limbic network
SMN	Sensorimotor network
DAN	Dorsal attention network
VN	Visual network

References

Anderson, M. (2014). *After phrenology: Neural reuse and the interactive brain*. Cambridge, MA: MIT Press.

Barlow, D., Frachione, T., Fairholme, C., Ellard, K., Boisseau, C., Allen, L., & Ehrenreich-May, J. (2011). *Unified protocol for transdiagnostic treatment of emotional disorders—therapist guide*. New York, NY: Oxford University Press.

Barrett, L. (2017). *How emotions are made: The secret life of the brain*. New York, NY: Houghton Mifflin Harcourt.

Barrett, L., & Satpute, A. (2013). Large-scale brain networks in affective and social neuroscience: Towards an integrative functional architecture of the brain. *Current Opinion in Neurobiology*, *23*(3), 361–372. http://doi.org/10.1016/j.conb.2012.12.012

Bastos, A., Usrey, W., Adams, R., Mangun, G., Fries, P., & Friston, K. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711. http://doi.org/10.1016/j.neuron.2012.10.038

Bermudez-Rattoni, F., Forthman, D., Sanchez, M., Perez, J., & Garcia, J. (1988). Odor and taste aversions conditioned in anesthetized rats. *Behavioral Neuroscience*, *102*(5), 726–732. http://doi.org/10.1037/0735-7044.102.5.726

Beutel, M., Scheurich, V., Knebel, A., Michal, M., Wiltink, J., Graf-Morgenstern, M Tschan, R., . . . Subic-Wrana, C. (2013). Implementing panic-focused psychodynamic psychotherapy into clinical practice. *Canadian Journal of Psychiatry/Revue Canadienne de Psychiatrie*, *58*(6), 326–334.

Binder, J., Desai, R., Graves, W., & Conant, L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796. http://doi.org/10.1093/cercor/bhp055

Binder, J., Frost, J., Hammeke, T., Bellgowan, P., Rao, S., & Cox, R. (1999). Conceptual processing during the conscious resting state: A functional MRI study. *Journal of Cognitive Neuroscience*, *11*(1), 80–93. http://doi.org/10.1162/089892999563265

Bogacz, R. (2017). A tutorial on the free-energy framework for modelling perception and learning. *Journal of Mathematical Psychology*, *76*(Pt B), 198–211. http://doi.org/10.1016/j.jmp.2015.11.003

Buhle, J., Silvers, J., Wager, T., Lopez, R., Onyemekwu, C., Kober, H., . . . Ochsner, K. (2014). Cognitive reappraisal of emotion: A meta-analysis of human neuroimaging studies. *Cerebral Cortex*, *24*(11), 2981–2990. http://doi.org/10.1093/cercor/bht154

Burešová, O., & Bureš, J. (1977). The effect of anesthesia on acquisition and extinction of conditioned taste aversion. *Behavioral Biology*, *20*(1), 41–50. http://doi.org/10.1016/S0091-6773(77)90473-4

Celeghin, A., de Gelder, B., & Tamietto, M. (2015). From affective blindsight to emotional consciousness. *Consciousness and Cognition*, *36*, 414–425. http://doi.org/10.1016/j.concog.2015.05.007

Chamberlain, S., & Robbins, T. (2013). Noradrenergic modulation of cognition: Therapeutic implications. *Journal of Psychopharmacology*, *27*(8), 694–718. http://doi.org/10.1177/0269881113480988

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. New York, NY: Oxford University Press.

Cools, R., Nakamura, K., & Daw, N. (2011). Serotonin and dopamine: Unifying affective, activational, and decision functions. *Neuropsychopharmacology*, *36*(1), 98–113. http://doi.org/10.1038/npp.2010.121

Daw, N., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. http://doi.org/10.1038/nn1560

Dehaene, S. (2014). *Consciousness and the brain*. New York, NY: Viking Press.

Dehaene, S., Changeux, J., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, *10*(5), 204–211. http://doi.org/10.1016/j.tics.2006.03.007

Dehaene, S., Charles, L., King, J.-R., & Marti, S. (2014). Toward a computational theory of conscious processing. *Current Opinion in Neurobiology*, *25*, 76–84. http://doi.org/10.1016/j.conb.2013.12.005

Feldman, H., & Friston, K. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, 215. http://doi.org/10.3389/fnhum.2010.00215

Foa, E., Hembree, E., & Rothbaum, B. (2007). *Prolonged exposure therapy for PTSD: Emotional processing of traumatic experiences therapist guide*. New York, NY: Oxford University Press.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–138. http://doi.org/10.1038/nrn2787

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O Doherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neuroscience and Biobehavioral Reviews*, *68*, 862–879. http://doi.org/10.1016/j.neubiorev.2016.06.022

Friston, K., Stephan, K., Montague, R., & Dolan, R. (2014). Computational psychiatry: The brain as a phantastic organ. *Lancet. Psychiatry*, *1*(2), 148–158. http://doi.org/10.1016/S2215-0366(14)70275-5

Gershman, S., Norman, K., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, *5*, 43–50. http://doi.org/10.1016/J.COBEHA.2015.07.007

Greenberg, L. (2010). *Emotion-focused therapy: Theory and practice*. Washington, DC: APA Press.

Greve, A., Cooper, E., Kaula, A., Anderson, M., & Henson, R. (2017). Does prediction error drive one-shot declarative learning? *Journal of Memory and Language*, *94*, 149–165. http://doi.org/10.1016/j.jml.2016.11.001

Hayes, S., & Smith, S. (2005). *Get out of your mind and into your life: The new acceptance and commitment therapy*. Oakland, CA: New Harbinger.

Henson, R., & Gagnepain, P. (2010). Predictive, interactive multiple memory systems. *Hippocampus*, *20*(11), 1315–1326. http://doi.org/10.1002/hipo.20857

Hermundstad, A., Bassett, D., Brown, K., Aminoff, E., Clewett, D., Freeman, S., … Carlson, J. (2013). Structural foundations of resting-state and task-based functional connectivity in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 6169–6174. http://doi.org/10.1073/pnas.1219562110

Hohwy, J. (2014). *The predictive mind*. New York, NY: Oxford University Press.

Kashdan, T., Barrett, L., & McKnight, P. (2015). Unpacking emotion differentiation: Transforming unpleasant experience by perceiving distinctions in negativity. *Current Directions in Psychological Science*, *24*(1), 10–16. http://doi.org/10.1177/0963721414550708

Kihlstrom, J., Mulvaney, S., Tobias, B., & Tobis, I. (2000). The emotional unconscious. In E. Eich, J. Kihlstrom, G. Bower, J. Forgas, & P. Niedenthal (Eds.), *Cognition and emotion* (pp. 30–86). New York, NY: Oxford University Press.

Lane, R., & Garfield, D. (2005). Becoming aware of feelings: Integration of cognitive-developmental, neuroscientific, and psychoanalytic perspectives. *Neuropsychoanalysis*, *7*, 5–30.

Lane, R., Ryan, L., Nadel, L., & Greenberg, L. (2015). Memory reconsolidation, emotional arousal and the process of change in psychotherapy: New insights from brain science. *Behavioral and Brain Sciences*, *38*, e1.

Lane, R., & Schwartz, G. (1987). Levels of emotional awareness: A cognitive-developmental theory and its application to psychopathology. *American Journal of Psychiatry*, *144*, 133–143.

Lane, R., Weihs, K., Herring, A., Hishaw, A., & Smith, R. (2015). Affective agnosia: Expansion of the alexithymia construct and a new opportunity to integrate and extend Freud's legacy. *Neuroscience & Biobehavioral Reviews*, *55*, 594–611. http://doi.org/10.1016/j.neubiorev.2015.06.007

Li, W., Zinbarg, R., Boehm, S., & Paller, K. (2008). Neural and behavioral evidence for affective priming from unconsciously perceived emotional facial expressions and the influence of trait anxiety. *Journal of Cognitive Neuroscience*, *20*(1), 95–107. http://doi.org/10.1162/jocn.2008.20006

Millner, J., & Palfai, T. (1975). Metrazol impairs conditioned aversion produced by LiCl: A time dependent effect. *Pharmacology Biochemistry and Behavior*, *3*(2), 201–204. http://doi.org/10.1016/0091-3057(75)90149-5

Monahan, J., Murphy, S., & Zajonc, R. (2000). Subliminal mere exposure: Specific, general, and diffuse effects. *Psychological Science*, *11*(6), 462–466. http://doi.org/10.1111/1467-9280.00289

Pang, R., Turndorf, H., & Quartermain, D. (1996). Pavlovian fear conditioning in mice anesthetized with halothane. *Physiology & Behavior*, *59*(4), 873–875. http://doi.org/10.1016/0031-9384(95)02137-X

Panksepp, J., Lane, R., Solms, M., & Smith, R. (2017). Reconciling cognitive and affective neuroscience perspectives on the brain basis of emotional experience. *Neuroscience & Biobehavioral Reviews*, *76*, *Part B*, 187–215. http://doi.org/10.1016/j.neubiorev.2016.09.010

Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, *134*, 17–35. http://doi.org/10.1016/j.pneurobio.2015.09.001

Reber, P. (2013). The neural basis of implicit learning and memory: A review of neuropsychological and neuroimaging research. *Neuropsychologia*, *51*(10), 2026–2042. http://doi.org/10.1016/j.neuropsychologia.2013.06.019

Roll, D., & Smith, J. (1972). Conditioned taste aversion in anesthetized rats. In M. Seligman & J. Hager (Eds.), *Biological boundaries of learning* (pp. 98–102). New York, NY: Appleton-Century-Crofts.

Rozin, P., & Ree, P. (1972). Long extension of effective CS-US interval by anesthesia between CS and US. *Journal of Comparative and Physiological Psychology*, *80*(1), 43–48. http://doi.org/10.1037/h0032831

Scherer, K. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion*, *23*(7), 1307–1351. http://doi.org/10.1080/02699930902928969

Sevenster, D., Beckers, T., & Kindt, M. (2014). Prediction error demarcates the transition from retrieval, to reconsolidation, to new learning. *Learning & Memory*, *21*(11), 580–584. http://doi.org/10.1101/LM.035493.114

Shiota, M., & Kalat, J. (2012). *Emotion* (2nd ed.). Belmont, CA: Cengage Learning.

Siemer, M., Mauss, I., & Gross, J. (2007). Same situation—different emotions: How appraisals shape our emotions. *Emotion*, *7*(3), 592–600. http://doi.org/10.1037/1528-3542.7.3.592

Smith, R. (2016). The relationship between consciousness, understanding, and rationality. *Philosophical Psychology*, 1–15. http://doi.org/10.1080/09515089.2016.1172700

Smith, R. (2017). A neuro-cognitive defense of the unified self. *Consciousness and Cognition*, *48*, 21–39. http://doi.org/10.1016/j.concog.2016.10.007

Smith, R., Akozei, A., Killgore, W., & Lane, R. (2018). Nested positive feedback loops in the maintenance of major depression: An integration and extension of previous models. *Brain, Behavior, and Immunity*, *67*, 374–397. http://doi.org/10.1016/j.bbi.2017.09.011

Smith, R., Killgore, W., & Lane, R. (2017). The structure of emotional experience and its relation to trait emotional awareness: A theoretical review. *Emotion*. http://doi.org/10.1037/emo0000376

Smith, R., & Lane, R. (2015). The neural basis of one's own conscious and unconscious emotional states. *Neuroscience & Biobehavioral Reviews*, *57*, 1–29. http://doi.org/10.1016/j.neubiorev.2015.08.003

Smith, R., & Lane, R. (2016). Unconscious emotion: A cognitive neuroscientific perspective. *Neuroscience and Biobehavioral Reviews*, *69*, 216–238. http://doi.org/10.1016/j.neubiorev.2016.08.013

Smith, R., Parr, T., & Friston, K. J. (2019). Simulating emotions: An active inference model of emotional state inference and emotion concept learning. bioRxiv 640813. https://doi.org/10.1101/640813

Smith, R., Lane, R., Parr, T., & Friston, K. (2019). Neurocomputational mechanisms underlying emotional awareness: insights afforded by deep active inference and their potential clinical relevance. *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1101/681288

Smith, R., Thayer, J., Khalsa, S., & Lane, R. (2017). The hierarchical basis of neurovisceral integration. *Neuroscience & Biobehavioral Reviews*, *75*, 274–296. http://doi.org/10.1016/j.neubiorev.2017.02.003

Smith, S., Fox, P., Miller, K., Glahn, D., Fox, P., Mackay, C., . . . Beckmann, C. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(31), 13040–13045. http://doi.org/10.1073/pnas.0905267106

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. London, England: MIT Press.

Tamietto, M., & de Gelder, B. (2010). Neural bases of the non-conscious perception of emotional signals. *Nature Reviews Neuroscience*, *11*(10), 697–709. http://doi.org/10.1038/nrn2889

Teigen, K. (1994). Yerkes-Dodson: A law for all seasons. *Theory and Psychology*, *4*, 525–547.

van den Heuvel, M., & Sporns, O. (2013). An anatomical substrate for integration among functional networks in human cortex. *Journal of Neuroscience*, *33*(36), 14489–14500. http://doi.org/10.1523/JNEUROSCI.2128-13.2013

Wilson, R., Takahashi, Y., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, *81*(2), 267–279. http://doi.org/10.1016/j.neuron.2013.11.005

Winkielman, P., & Berridge, K. (2004). Unconscious emotion. *Current Directions in Psychological Science*, *13*(3), 120–123. http://doi.org/10.1111/j.0963-7214.2004.00288.x

Winkielman, P., Berridge, K., & Wilbarger, J. (2005). Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Personality & Social Psychology Bulletin*, *31*(1), 121–135. http://doi.org/10.1177/0146167204271309

Winkielman, P., Zajonc, R., & Schwarz, N. (1997). Subliminal affective priming resists attributional interventions. *Cognition & Emotion*, *11*(4), 433–465. http://doi.org/10.1080/026999397379872

Yeo, B., Krienen, F., Sepulcre, J., Sabuncu, M., Lashkari, D., Hollinshead, M., . . . Buckner, R. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*(3), 1125–1165. http://doi.org/10.1152/jn.00338.2011

Zemack-Rugar, Y., Bettman, J., & Fitzsimons, G. (2007). The effects of nonconsciously priming emotion concepts on behavior. *Journal of Personality and Social Psychology*, *93*(6), 927–939. http://doi.org/10.1037/0022-3514.93.6.927