On Valence: Imperative or Representation of Value?

Peter Carruthers

Affective valence is increasingly thought to be the common currency underlying all forms of intuitive, non-discursive decision making, in both humans and other animals. And it is thought to constitute the good or bad (pleasant or unpleasant) aspects of all desires, emotions, and moods. This article contrasts two theories of valence. According to one, valence is an experience-directed imperative ('more of this!' or 'less of this!'); according to the other, valence is a representation of adaptive value or disvalue. The latter is shown to provide the better account.

1. The Two Theories

Affective mental states are a broad class. They comprise bodily states of pain, tiredness, and discomfort, as well as bodily pleasures like drinking when thirsty, pleasant touch, and orgasm. But they also include emotional states of grief, fear, anger, joy, and so on; felt desires and repulsions; as well as moods like anxiety, depression, and cheerfulness. Moreover, all affective states have a valence component in common (whether positive or negative). And it is increasingly accepted among cognitive scientists who study the nature of affect that valence is the common currency of non-discursive, intuitive forms of decision making, at least (Gilbert and Wilson [2005], [2007]; Levy and Glimcher [2012]; Bartra et al. [2013]; Seligman et al. [2013]). Or perhaps better, valence is a common scale, rather than a common currency (Rolls [2015]). This is because, while positive and negative valences can be compared with one another for extent and traded off against one another, they do not exist on a single representational dimension organized around a midpoint. Rather, the networks that process positive and negative valence in the brain are largely distinct from one another, as we will see—suggesting that the two forms of valence comprise two different (albeit directly comparable) types of analogue-magnitude representation.

Recently, two competing accounts of the nature of valence have been proposed and defended by philosophers: evaluativism (Carruthers [2018]; Cochrane [2019]) and imperativism (Barlassina and Hayward [2019a], [2019b]). According to evaluativism,

the two valences are analogue-magnitude representations of value (non-conceptual goodness and badness, respectively). According to imperativism, in contrast, the valences are graded-strength imperatives with the content, 'more of this!' and 'less of this!'. The goal of the present article is to offer a comprehensive comparison of the two approaches, demonstrating the superiority of evaluativism. At stake in the debate is nothing less than how we should understand the representations that lie at the core of non-discursive decision making, in both humans and nonhuman animals.

The debate over the nature of valence grows out of what is now a large philosophical literature on the painfulness of pain.[1] Here, too, two broad camps are pitted against one another. Everyone in this debate accepts that pain experiences are composed of a pair of dissociable dimensions: there is a sensory component, which represents where tissue-damage is located on or within the body (or perhaps better, risk of tissue damage; Bain [2017]); and there is a negative valence component, which makes pain experiences painful or bad. But then in one camp are those who think that the negative valence of pain is a non-conceptual representation of badness (Cutter and Tye [2011], [2014]; Bain [2013], [2019]), and in the other camp are people who think that the painfulness of pain is composed of an imperative-like content, 'less of this!'—where the referent of 'this' can either be the sensory component of pain (Martínez [2011], [2015]) or the whole experience composed of the sensory component together with the imperative content itself (in which case the imperative really takes a reflexive form, 'less of me!'; Barlassina and Hayward [2019b]).[2]

The two forms of imperativism pretty nearly coincide in their treatment of pain. The non-reflexive theory claims that the painfulness of pain is composed of a 'less-of' imperative targeting just the sensory component of pain. The reflexive account differs only in claiming that the target of the imperative is not just the sensory component, but the entire experience comprising both sensation and valence. The differences remain small when the two forms of imperativism are generalized to other affective states, provided that the indexical in the non-reflexive version is thought to pick out the experience into which it is bound. Consider the negative valence component of the fear one feels when unexpectedly confronting a grizzly bear in the forest. On a non-reflexive account, the 'this' in, 'less of this!' refers to the experience of seeing the bear approach; whereas for the reflexive view it refers to the experience of

---

[1] I should emphasize that although debates among philosophers about the nature of valence (as well as about the painfulness of pain) focus mostly on the human case, much of what we know about affective states in general, and valence in particular, derives from studies conducted with other animals (Panksepp [1998]; Izard [2007]; Berridge and Kringelbach [2013]).

[2] There are other types of view in addition to these two, of course. These could include desire-based accounts of pain (Heathwood [2007]) and psychofunctionalist theories (Aydede [2014]). But to the best of my knowledge, no attempt has been made to develop these alternative theories into accounts of positive and negative valence in general. The only fully general philosophical theories of valence at present are evaluativism and imperativism. Note, too, that although Klein ([2015]) offers an imperative theory of pain, it is an imperative theory of the sensory component of pain. Klein is explicit that he lacks a theory of the badness, or painfulness, of pain.

seeing the bear approach together with the imperative component itself. The difference is a subtle one, and will not matter for our purposes.

The differences are much more substantial, however, if the non-reflexive from of imperativism is generalized in a different way, and the imperative component is taken to target items in the world. Martínez ([2011]) tentatively proposes a view of this sort. On this view, the negative valence component of fear of an approaching bear would be composed of an urge to get away, which we can represent as 'less of this!' (meaning the approach of the bear, rather than the experience of an approaching bear). But as we will see in section 2, all affective states are thought to involve urges to act that are quite separate from the valence component of those states. So a world-directed form of imperativism would pick on the wrong component of affect to identify with valence.

If evaluativism and experience-directed imperativism turn out to be the only two games in town, as accounts of the valence component of all affective states, then the stakes are high. For if, as some believe, valence forms the basis for all decision making (both discursive and non-discursive; Carruthers [2015], [2018]), then experience-directed imperativism entails a form of motivational hedonism. All decisions will be taken in order to secure specific sorts of experience for oneself. This may have profound implications for our conception of ourselves as moral agents, and perhaps even for the very existence of virtues of character.

Before we compare imperativism in more detail with evaluativism, I propose, in section 2, to sketch some of what is known from the science of affect. This is because all parties to this debate are naturalists, who take themselves to be offering the best theory of what valence really is. They are not analysing our ordinary concepts, nor elucidating our folk-psychological beliefs. So it is a strong constraint on theorizing in this domain that the theories should be at least consistent with the current state of scientific knowledge—or better still, they should aim to provide the best unifying interpretative account of that knowledge.

2. The Science of Affect

Despite ongoing disputes about whether, for example, there are innate discrete emotions or just an initially undifferentiated structure of valence and arousal dimensions (Ekman [1992]; Barrett et al. [2007]; Izard [2007]; Panksepp and Watt [2011]), there is pretty general consensus in the field about the properties that most affective states share. In fact, affective states form a natural kind in something like the sense of Boyd ([1991], [1999])—they constitute a homeostatic property cluster.

First, most affective states result from appraisals of the value of some property, thing, or event, whether currently perceived, or merely remembered or imagined (Moors et al. [2013]). In some instances (for example, fear), the appraisal mechanisms are entirely world-focused (for example, an approaching bear), whereas in others (for example, hunger and thirst) they also take input from the current state of

the body (for example, calorie depletion or dehydration), and in yet others (for example, pain) their input is almost entirely from the body (Ashar et al. [2017]). Even free-floating moods such as depression or cheerfulness are thought by some to reflect more generalized appraisals of the opportunities afforded by the current environment (Eldar et al. [2016]). As should be clear from these examples, appraisal processes need not be judgement-like (let alone conscious). They can be associative in nature, and are sometimes 'hard-wired' and uncontrollable (as when suddenly looming objects initiate a fear response).[3]

Second, the appraisals that initiate many affective states automatically activate a suite of bodily changes or 'arousal'—including increases in heart rate and breathing rate, preparatory release of cortisol and adrenaline into the bloodstream, increased sweating, and so on. Many of these changes are initiated before one even becomes conscious of the stimulus (Adolphs and Anderson [2018]). In fact, most affective states can be organized according to level of arousal, from depression and contentment at one end (low arousal) to joy, rage, and abject terror at the other (high arousal; Russell [2003]).

Third, most affective states directly, and automatically, activate motor plans of various sorts, depending on the nature of the initial appraisal (LeDoux [2012]). These motor tendencies will need to be actively suppressed if they are not to be carried through to completion. In the case of emotions, they include motor instructions for facial expressions (the fear face, the anger face, the surprise face, and so on) as well as specific sorts of bodily behaviour (running or freezing in the case of fear, depending on the affordances of the situation; approaching and looking more closely, in the case of curiosity). Most felt desires, too, at least initiate basic approach or retreat motor plans, as we can tell from reaction-time studies involving movement forwards or backwards using positive and negative stimuli (van Dantzig et al. [2008]; Krieglmeyer et al. [2010]). Pain will directly activate motor plans for nursing and protecting the painful part. And even moods issue in characteristic overall bodily postures, muscle tone, and modes of activity (slumped lassitude, in the case of depression; energetic gait and upright posture, in the case of cheerfulness).

It should be emphasized that these action tendencies are produced directly and automatically, and independently of the valence component of affective states. That is to say: the valence component and the action tendencies have a common cause (the appraisal), rather than either one proceeding via the other (LeDoux [2012]).[4] This is what causes the deep problem for any first-order version of imperativism, noted above: for these action imperatives are distinct from the valence signals that

---

[3] Note that I am here using the term 'appraisal' somewhat more broadly than is common within so-called appraisal theories of emotion (Scherer [2009]), where appraisals often seem to be conceived as judgement-like.

[4] This is not to say that valence and action do not subsequently interact, of course. On the contrary, valenced feedback will strengthen action tendencies over time, issuing in habits. And urges to act will create negative valence when suppressed, since self-control is generally experienced as effortful (Shenhav et al. [2017]).

the world-directed theory tries to identify them with. (A valence 'signal' is the mental state or representation that carries the valenced content.)

One way of thinking of experience-directed imperativism, in contrast, is that it treats valence as a wholly unspecific generalization of the automatic action-initiation tendencies produced by all affective states. On this account, valence is composed of an instruction to do something (unspecified) to extend the affective experience of which it forms a part (in the case of positive valence) or to bring that experience to an end (in the case of negative valence). While fear automatically issues in the motor imperative 'get away from that [approaching bear]!', the valence component of fear can be seen as the more general imperative 'do something to get less of this [fear] experience!'.

Fourth, affective states have a valence component. Valence is thought to play at least three role types:

(1) When bound into the contents of perception, imagination, or memory, valence signals play a critical role in decision making. In prospective reasoning, especially, people imaginatively envisage alternative actions and outcomes, responding affectively to each. In the absence of any conflict with existing goals, the alternative that ranks highest in overall valence is generally chosen (Gilbert and Wilson [2005], [2007]; Levy and Glimcher [2012]; Seligman et al. [2013]). Moreover, since birds, too, can engage in multi-step planning (Taylor et al. [2010]; von Bayern et al. [2018]), and can do so even though the relevant components for the decision are occluded (Gruber et al. [2019]), it seems that prospection is by no means proprietary to humans.

(2) Valence signals play a critical role in evaluative learning. When the positive valence created by an outcome is higher than expected, an error signal is sent to the underlying appraisal system to ramp up the positive value attached to outcomes of that sort. And when the positive valence created is lower than expected, the value attached to outcomes of that kind is incrementally lowered (Schultz et al. [1997]). Something similar happens with the learning of negative values (Pfabigan et al. [2011]). Importantly, although affective learning is enhanced by attention, valenced stimuli do not have to be conscious for affective learning to take place (Kirk-Smith et al. [1983]; Greenwald and De Houwer [2017]).

(3) Valence signals also operate widely throughout cognition, especially within perceptual systems where they bias processing toward positive or negative stimuli (Barrett and Bar [2009]; Persichetti et al. [2015]), and within the bottom-up salience or 'relevance' attentional system, where they influence the competition for top-down attentional resources and entry into consciousness (Corbetta et al. [2008]). Note that it follows straightforwardly from this latter role that valence itself need not be conscious, since it can be the valence

that attaches to an unconscious state that subsequently causes it to become conscious.

One other important fact about the valence component of all affective states is that it can be influenced (ramped up or down) by evaluative beliefs and concepts, as well as by low-level associations. Thus, expecting a wine to taste good makes it taste better (Plassmann et al. [2008]), expecting something to smell like body odor makes it smell worse (de Araujo et al. [2005]), expecting something to be painful makes it more so (the nocebo effect), and expecting something to be less painful makes it less so (a placebo effect; Ashar et al. [2017]). It is thus a mistake to think of affective processing as entirely 'bottom-up'. On the contrary, although bottom-up appraisals of a stimulus or memory image will initiate downstream effects on behaviour, arousal, and valence swiftly and automatically, later processes of conceptualization and reappraisal can (and often do) modify or modulate those outputs thereafter. For example, the experience of leaping from an airplane can be terrifying or merely exhilarating, depending on one's belief that one is wearing an effective parachute.

3. Imperativism: An Initial Look

The forms of imperativism now under discussion are experience-directed ones, especially (for concreteness) the reflexive version proposed and defended by Barlassina and Hayward ([2019a], [2019b]) and Barlassina ([2020]). On this view, positive valence is an instruction of the form 'more of me!' (referring to the experience of which the valence is a part), whereas negative valence is an instruction of the form 'less of me!' (again, referring to the overall experience in question). These imperatives can be of greater or lesser strength, or urgency, of course, allowing for comparison across cases.

One advertised advantage of the view is that it explains how valence can be directly motivating. For on this account valence is, in effect, a state of initiating a search to do something (unspecified) to increase or extend (in the case of positive valence), or decrease or shorten (in the case of negative valence), the experience of which it is a part. Valence is an urge—of greater or lesser magnitude—towards unspecified action with a particular experiential outcome. Since positive valence, for an imperativist, is effectively a high-level motor instruction ('do something to get more of me!'), its impact on action is a direct one—with facilitators, perhaps (for example, beliefs about what might achieve 'more of me') but no intermediaries.

For an evaluativist, in contrast, the motivational role of valence is indirect, operating via one (or both) of two routes. First, it can have an immediate impact on decision-making processes, which in turn result in intentions to act. It is those intentions that motivate action directly. And second (as Cochrane [2019] argues), it can attract attention toward the evaluated target, strengthening existing action tendencies that have been caused by the initial appraisal. But the question here really is: why should we

think that valence is directly rather than indirectly motivating? For both sides can agree that in cases where there are decisions to be made, the main effect of valence on subsequent action is via its effects on decision making. Whether the latter process is best thought of as competitive selection among indicators of value to form an intention to act, or competition among imperatives leading to selection of an action (which is presumably then intended), seems not to provide a strong advantage either way.

An initial concern one might have about experience-directed forms of imperativism arises from their meta-representational nature. For notice that valence, on such accounts, involves a reference to the experience of which it is a part, and experiences are representational states. (Moreover, even if one denies that experience is representational, valence must at least be meta-mental. For everyone agrees that experiences are mental states.) Barlassina and Hayward seemingly do not notice this implication of their theory, perhaps because they refer to it as 'same-order' rather than 'higher-order'. (In this they follow Kriegel's ([2007]) same-order theory of consciousness.) But the fact remains that the reflexive imperative refers to the overall experience of which it is a part.

Why might this be a concern? Well, it is much disputed whether any creatures outside of humans are capable of meta-representation, or of representing their own or others' mental states (Penn et al. [2008]; Smith et al. [2014]; Heyes [2015]; Carruthers and Williams [2019]). But valence is thought to be common to all organisms that are capable of evaluative learning, which incorporates almost all living creatures, including many insects and other invertebrates, including snails (Kobayashi et al. [1998]). So it might seem that experience-directed forms of imperativism are burdened with a highly controversial commitment.

Although this implication of experience-directed imperativism may be surprising, it perhaps does not present any deep difficulty for the approach. For notice that valence, on this view, refers to the experience of which it is a part via a pure indexical. So it could do its work in the absence of anything resembling a 'theory of mind', and could do so in creatures that lack mental-state concepts of any sort. Moreover, others, too, have postulated that forms of purely non-conceptual metacognition might be widespread across the animal kingdom. Thus Shea ([2014]), for example, argues at length that the error signals involved in evaluative learning have meta-representational contents, as well as arguing (at lesser length) that the same is true of the error signals involved in the online control of action. In fact, the existence of non-conceptual forms of metacognition need have no bearing on existing debates about non-human meta-representation in general, which has almost exclusively concerned the presence or absence of concept-like representations of mental states in oneself or other creatures.[5]

---

[5] Some have also postulated forms of implicit, or procedural, metacognition, that can likewise occur independently of meta-representational symbols or mental-state concepts (Dokic [2012]; Proust [2014]). This, too, need not be controversial. For example, the salience systems that compete to control top-down attention have implicit metacognitive functions, since they serve to modulate and control the contents

With that said, just a little further reflection suggests that experience-directed imperativism is at best seriously incomplete. It needs at least to be expanded into the claim that valence is a form of what Millikan ([1995]) calls a 'pushmi-pullyu' representation, having indicative as well as imperative content. This is because valence reliably carries information about the adaptive value of items or events in the environment or body, and because the presence of such information explains how the role of valence has become stabilized in decision making and evaluative learning. So, given our best accounts of representational content in cognitive science (whether informational or teleosemantic in nature; Dretske [1988]; Millikan [1989]; Rupert [2018]; Shea [2018]), valence is a representation of adaptive value (in addition to having imperative content, perhaps). But all this will take some unpacking.

As we noted in section 2, affective states are initiated by appraisals of stimuli, events, or situations. But to what properties are appraisals ultimately sensitive? What are they designed to track and respond to? The answer is implicit in long-standing theories of affective learning. Affective states are produced by the presence of, or in anticipation of, both primary and secondary reinforcers and punishers (Grabenhorst and Rolls [2011]; Juechems and Summerfield [2019]). Primary reinforcers and punishers are things whose adaptive value or disvalue have been fixed through natural selection. Primary reinforcers include such things as eating when hungry, drinking when thirsty, orgasm, and (in humans) indications of liking or admiration from community members. Primary punishers include tissue damage or risk of tissue damage, sickness/nausea, signs of immediate physical danger, and (among humans) indications of hatred or disrespect from community members. All primary reinforcers and punishers have played a critical role in the inclusive fitness of individual humans and ancestral organisms over significant stretches of evolutionary time, and their value or disvalue is now innately fixed (albeit modifiable through subsequent learning).

Secondary reinforcers or punishers are things or properties that have acquired their value or disvalue though affective learning. They are things or properties that have become associated with, or predictive of, primary reinforcers or punishers. For a clearcut example, consider the Garcia effect (Garcia et al. [1955]). When a mammal has become seriously sick with nausea and vomiting, an appraisal process from a search of memory identifies the most likely cause among items recently eaten. One-shot learning then results in that item evoking disgust (becoming a secondary punisher) for some significant time into the future. This will be familiar to most readers: you wake up vomiting one morning having eaten a chicken salad the night before, and thereafter (often for weeks afterwards) you cannot bear the thought of eating chicken again. If someone forces you to eat chicken, you are likely to find it extremely unpleasant, feeling nauseated, and perhaps even vomiting again.

---

that enter working memory. Yet attentional mechanisms of this general sort are quite widespread in the animal kingdom, being common to all mammals and birds, and likely all vertebrates (Gruberg et al. [2006]; Knudsen [2011]; Mysore and Knudsen [2013]).

The appraisal processes that result in affective states, then, are sensitive to adaptively valuable and disvaluable properties of the world or body, and to properties that correlate with or are predictive of adaptively valuable or disvaluable properties. The positive valence signals that result from such appraisal processes therefore carry information about, and reliably indicate, an overall balance of valuable over disvaluable properties in the item or event appraised. (Of course, over time, one can often separate out the valuable from the disvaluable properties by switching attention selectively between them.) And likewise, negative valence signals reliably indicate an overall balance of disvaluable over valuable properties in that item or event.

Moreover, it is plainly the case that valence signals have the roles in human cognition that they do as a result of the information that they carry, and that the computations into which they enter influence behaviour because of their information-carrying role. Decision making that selects the most positively valenced option (or rejects and avoids the most negatively valenced) will be adaptive precisely because valence is tracking adaptive values and disvalues in the environment or the agent's own body. And the way in which error signals drive affective learning (resulting from valence that is higher or lower than expected) is adaptive, too, precisely because such signals provide an updated indication of current adaptive value and disvalue. Moreover, valenced contents attract attention because of their likely relevance to the organism's values.

On any viable version of informational or teleosemantic theory of representational content, then, it follows that valence is a representation of adaptive value or disvalue. More specifically, according to the theory developed by Shea ([2018]), valence will represent adaptive values because carrying that information explains how the roles played by the signals in question have become stabilized. It is because they carry information about adaptive values that the roles of valence signals are now as they are. And according to the account provided by Rupert ([2018]), valence will represent adaptive values because carrying that information explains why valence signals enter into the computations that guide decision making and action in the way that they do. Moreover, valence signals are employed by downstream consumer systems in the manner one would expect for representations of value. For they give rise to judgements of goodness and badness. The positive valence created by anticipation of eating a slice of chocolate cake will lead one to judge that eating the cake would be good. And the negative valence created by an approaching grizzly might lead one to exclaim, 'Oh, this is bad!'

So, to repeat, imperativists about valence must at least accept that valence is a pushmi-pullyu representation. It represents the presence of adaptive value or disvalue, in addition to demanding more or less of the experience in which it is embedded. The burden is then placed on imperativists to show that it is not sufficient to treat valence as merely a representation of value or disvalue. They need arguments for claiming that valence also has imperative content. Otherwise evaluativism gets to win by default. Imperativists have not yet taken up this challenge. Instead, for the most

part they have argued negatively, offering counterexamples to evaluativism. These will be discussed in section 5. What we consider next is how successfully imperativism can accommodate scientific findings about valence.

4. Imperativism and the Science of Valence

How well can imperativism accommodate what cognitive science tells us about valence and its roles? Consider, first, the role of valence in evaluative learning, sketched in section 2. When an event of a familiar type is anticipated (for example, one is about to eat an apple), an appraisal process accesses stored values for events of that type and forms an expectation of the likely degree of valence to be created by it (in this case, for most people, eating the apple would be expected to produce positive valence, or pleasure). The event then occurs (one eats the apple) and the degree of valence turns out to be higher, or lower, than anticipated, resulting in the stored value for events of that kind being ratcheted up or down by some increment.

Once one recognizes that appraisal processes are sensitive to indicators of adaptive value (as argued in section 3), one has to acknowledge that what is stored, as a result of evaluative learning, is at least a representation of value for the kind of event in question (in this case, eating apples). And when error signals are received, it is the stored value attaching to events of that type that gets changed. Now it might well be, as imperativists claim, that stored values are also stored imperatives. That is, when eating an apple is anticipated, an expectation also gets formed for a particular intensity or urgency of an imperative 'more of me!' to be experienced when the apple is eaten. And then when one eats the apple, one might feel an urge to continue the experience of eating that has greater or lesser urgency than expected, incrementally changing not only the stored value attached to apple-eating but also the stored expectation for future intensities of experience-directed imperatives.

It appears that imperativists can offer an interpretation that is consistent with the science of evaluative learning. But it also seems that the interpretation in question is unnecessary. Given that appraisal processes are sensitive to, and store expectations of, indicators of adaptive value; and given that valence, too, carries information about, and represents, degrees of value; then the entire story of affective learning can be told in terms of value. Appraisal systems store values and anticipate experiences of value; and then when the anticipated events happen, they cause degrees of experienced value (valence) that are higher or lower than expected, issuing in incremental change. There is simply no need to appeal to imperative content.

Something similar can be said about the role of valence in prospective decision making. Cognitive neuroscientists who study how decision making is realized in human and animal brains, and who develop formal models of the processes involved (often described as 'neuroeconomics'), assume that three sorts of information get integrated. When one is choosing among options or outcomes, one integrates the values one attaches to those outcomes with the likelihood of achieving them, as well as

with the energetic costs of acting, to select the best overall option (Botvinick et al. [2009]; Levy and Glimcher [2012]; Winstanley and Floresco [2016]). Valence figures in this account twice over: once in the form of the signals representing the value one attaches to each outcome, and once in the signals of expected effort (negative valence) of acting. Since these models are generally expressed in terms of computations of value, rather than computations that integrate likelihoods with positive and negative imperatives of various strengths, it seems that some form of evaluativism is implicit in the science.[6]

Now admittedly, as Barlassina and Hayward ([2019a]) point out, some of the social scientists who study prospection and prospective reasoning in humans express their theories using the language of anticipated positive or negative experience (Gilbert and Wilson [2005], [2007]; Miloyan and Suddendorf [2015]). This way of describing their findings suggests that valence is experience-directed rather than world-directed; and the only representational theory that fits the bill is one or other kind of experience-directed imperativism. In fact, however, the experience-directed aspect of these psychologists' views does no serious work in their theories. It may amount to little more than the sort of naïve hedonism about motivation and choice that has always been perennially tempting for people. It may merely be a hedonist gloss on what these psychologists are really interested in studying, which has to do with the various ways in which anticipations of value or hedonic experience can be biased and inaccurate. Whether the prospection errors that have been discovered and charted in this literature are errors in predicted value or errors in the predicted strength of experience-directed imperatives has no bearing on the discoveries themselves.

Thus far in this section I have proceeded under the assumption that valence at least represents degrees of adaptive value. (This claim was defended in section 3.) Given that this is true, it seems that there is no scientific need to appeal, in addition, to imperative content. Rather than being a sort of pushmi-pullyu representation, the content of valence can be treated throughout as purely a representation of value. There is an additional, more direct, argument against imperativism arising out of the science of affect, however, which does not depend on the assumption that valence represents value. This is because imperativism creates predictions about where, in the brain, signals of valence should be found. But these predictions turn out to be false.

One would expect that imperative content would be realized in networks closely related to those that initiate and control motor actions. For as we noted earlier, the imperative contents 'more of me!' and 'less of me!' amount to the instruction 'do something to get more/less of me!'. One might therefore expect it to be realized in some of the more abstract (less movement-specific) regions of premotor cortex. This is located in a band across the outside upper surface of prefrontal cortex, immediately

---

[6] And for good reason, one might think. For cognitive neuroscientists are well aware that valence signals carry information about adaptive values, of course; and they are—arguably—at least implicitly committed to some version of informational or teleosemantic theory of mental representation (Rupert [2018]; Shea [2018]).

anterior to motor cortex (which in turn lies just forward of the central sulcus that separates frontal from parietal cortices). Alternatively, one might think of imperative content as being tantamount to an active goal to sustain or get more of (or eliminate or have less of) the experience in question. This leads to a similar prediction, since the networks for prospective memory ('what am I going to do?') and for active goal maintenance are centred on dorsolateral prefrontal cortex, which overlaps heavily with the regions just mentioned (Momennejad and Haynes [2012]; Cona et al. [2015]).

Where are valence signals found in the brain, then? Not in the regions just described. Simplifying hugely, positive valence is realized in networks linking especially subcortical regions of the ventral striatum with regions of orbitofrontal cortex, which is located just behind the eyes, beneath the lower surface of prefrontal cortex; negative valence, in contrast, is mostly realized in networks linking the amygdala, anterior insula (hidden in the fold between temporal cortex and ventral prefrontal cortex), and a distinct set of regions of orbitofrontal cortex (Grabenhorst and Rolls [2011]; Yarkoni et al. [2011]). Orbitofrontal cortex is thought to code for outcome values alone, binding valence representations computed in interaction with subcortical structures with incoming sensory or imagistic representations; nearby regions of ventromedial and medial prefrontal cortex (on the lower inside surface of prefrontal cortex) are thought to integrate valence with likelihood and energetic costs to provide representations of overall expected value; and competition among options and with existing goals is resolved through interactions between anterior cingulate cortex (just above medial prefrontal cortex, immediately above the corpus callosum) and dorsolateral prefrontal cortex, issuing in decisions to act (Wallis [2007]; Shenhav et al. [2014], [2018]).[7]

The evidence from cognitive neuroscience, then (albeit highly simplified here) suggests that valence, on the one hand, and motivation, on the other, are realized in very different non-overlapping networks in the brain. It is conceivable that cognitive neuroscientists have gotten it badly wrong, and that orbitofrontal cortex, for example, is really a high-level motor area, coding for high-level motor instructions of the form 'more/less of me!' or 'do something to get more/less of me!'. But it would require very strong philosophical arguments to warrant such a wholesale reworking of current scientific understanding. Indeed, it is not clear that anything a philosopher could say, at this point, would (or should) be sufficient to overturn an established scientific consensus.[8]

Finally, in this discussion of how well imperativism can handle findings from the science of affect, consider the fact that positive and negative valence signals are thought

---

[7] It is worth noting that existing goals and intentions are among the important top-down influences on affective processing. Goal achievement, or indications of progress towards a goal, will issue in positive valence; whereas goal frustration, or indications of obstacles to achievement, will issue in negative valence. In fact, goals and intentions can helpfully be thought of as secondary reinforcers and punishers.

[8] Note that only a sample of recent papers has been cited above for what I claim to be the scientific consensus. These could easily have been supported by many, many, more. For book-length discussions, see (Rolls [2014]; Adolphs and Anderson [2018]).

to play an important role in the 'salience' attentional system, competing for the control of top-down attention (Corbetta et al. [2008]). Consider mind-wandering, or daydreaming. One is initially focused on a necessary but uninteresting task—proof-reading an article for publication, for example. In the meantime, both peripheral experiences and memories sparked by associative processes are monitored unconsciously. On reading the word 'island' one suddenly finds oneself off-task and flooded with images of tropical fish and banana groves. How does this happen? Memories of a previous vacation in the Caribbean are activated by association with the word, initially unconsciously. But they are appraised positively, and their positive valence is enough to attract attention, leading them to burst suddenly into consciousness and working memory, causing one to abandon one's primary task.

Now, imperativism can handle this sort of example easily enough. For the positive valence attaching to memories of a Caribbean island takes the form of the instruction 'more of me!'—referring to the experience of remembering one's vacation. And then the 'action' needed to get more of that experience is to direct top-down attentional signals at them, leading them to become conscious.

Matters are quite otherwise, however, when it comes to explaining the role of negative valence in attracting and sustaining attention. Think of how a gruesome accident scene will attract one's attention. People say, 'I couldn't help looking, it was so horrible'. The experience is strongly negatively valenced, so imperativists must say that the valence comprises the instruction 'less of me!'—referring to the experience of seeing the scene. This then predicts that one would automatically look away, or never even notice. But one does not, of course: most people are fascinated, despite how unpleasant it is. From an evaluativist perspective, in contrast, this is easy to explain. For adaptively bad things as well as adaptively good ones are highly relevant to know about, and so attract one's attention (Cochrane [2019]).

In summary of the argument so far, then, imperativism confronts a number of difficulties when brought into contact with affective neuroscience. The first is that it is unnecessary, given that valence appears to be at least a representation of adaptive value. The second is that it makes the wrong predictions about where, in the brain, we should expect valence signals mainly to be found. And the third is that it wrongly predicts that negative valence in the salience system should automatically direct attention away from highly negative events.

Evaluativism, in contrast, has a nice synergistic fit with the cognitive science of affect. As we noted in sections 2 and 3, affective states result from appraisals of the personal significance (the adaptive value or disvalue) of items and events in the environment or body. And according to evaluativism, one output of these appraisal processes (positive or negative valence) is itself an analogue-magnitude representation of the very same adaptive values and disvalues—it is not a representation of a wholly different sort ('more of this experience!'), as imperativism claims. Moreover, an evaluativist account of the content of valence fits well with the theories provided by cognitive neuroscientists of human decision-making processes and of the brain

networks involved, as well as with the mechanisms involved in affective learning through valence error signals, as we have seen in sections 3 and 4. Indeed, the entire account of the role of valence in cognition can be told in terms of analogue representations of value, without needing to mention anything imperative-like.[9]

5. Replies to Objections

As we noted in section 3, much of the case in support of imperativism comes in the form of arguments and counterexamples against its central rival, evaluativism, as well as in appeals to the conscious phenomenology of affective experience. I will first briefly consider the latter, before discussing the main alleged counterexamples.

It is far from clear that appeals to introspection can gain much traction in this debate. Both sides agree that valence is amodal (non-sensory), and that it can accompany, or be a component of, conscious experiences across all sensory modalities. According to imperativism, valence is felt as an urge to continue or discontinue those very experiences, disposing us to judge (conceptually) that the represented contents of those experiences are good or bad. According to evaluativism, in contrast, valence is a non-conceptual analogue-magnitude representation of the values attaching to those contents, again leading us to judge (conceptually) that those things are good or bad, and issuing in motivation to pursue or avoid them.[10] In both cases one finds oneself inclined to make judgements of the goodness or badness of worldly objects, and in both one finds oneself motivated to pursue or reject them. It is unclear that introspection affects the power to discriminate between the alternative theories. Nevertheless, Barlassina and Hayward ([2019a]; Barlassina [2020]) at least think that there are particular kinds of conscious affective experience that evaluativism has difficulty accounting for, as we now discuss.

Barlassina and Hayward ([2019a]) claim that pure mood states provide counterexamples to evaluativism. This is because in such cases there is (allegedly) no worldly object or event to be evaluated; so there is nothing here to seem (non-conceptually) bad. But although moods can lack an identifiable represented cause (their cause can rather be chemical in nature), they arguably focus on the world, nevertheless. The world in general is apt to seem flat and colourless to someone who is depressed (Lambie and Marcel [2002]). One looks up at the Mediterranean blue sky and thinks, 'Meh!'. One

---

[9] It may be worth emphasizing again, however, that this does not mean that there are no imperative-like representations embedded in affective states generally. On the contrary, as we noted in section 2, all affective states involve urges to act of various sorts that are directly caused by the appraisal process. But these urges are distinct from, and initially independent of, the valence component of affect.

[10] Valence does not represent goodness or badness as such, of course. It is more primitive than, and prior to, any concepts of goodness and badness. A useful comparison is with experiences of numerosity, or approximate number, which are also non-conceptual and amodal in nature. These exist prior to and independently of concepts of number, and are likewise common to many creatures besides ourselves (Dehaene [1997]; Jordan et al. [2008]; Izard et al. [2009]). One can have impressions of more or less in number, or of approximate numerical equality, independently of the deployment of the concepts of more, less, equal, or (of course) number. Likewise, one can have impressions of more or less goodness or badness, but without yet deploying any concepts of goodness or badness.

looks at a bank of azaleas in a riot of colour and thinks, 'Meh!'. One anticipates listening to one's favourite piece of music and thinks, 'Meh!'. Now, granted, failing to experience positive affect is not the same as experiencing affect that is negative; and yet depression is surely a negative affective state. But Barlassina and Hayward display insufficient awareness of the flexible nature of appraisal mechanisms, which can operate on higher-order as well as first-order contents, and can shift with the shifting direction of attention.

To illustrate: looking at a blue sky or an azalea in bloom, knowing that they normally evoke joy (for an evaluativist, normally seeming [non-conceptually] good), but failing to have any positive response, that fact itself can become an object of appraisal. It then seems (non-conceptually) bad that one fails to experience the usual joy. In effect, depression can be characterized as a state in which nothing seems good, and the fact that nothing seems good seems bad. Likewise, when one shifts attention to one's low-energy, listless, bodily state (characteristic of many forms of depression), that state, too, seems (non-conceptually) bad, especially in comparison with how one's body normally feels. Parallel things can then be said about positive moods, like cheerfulness—everything is apt to seem good (or better than normal), and the fact that everything seems good (and that one feels full of energy) itself seems good. The challenge for Barlassina and Hayward is to provide an example that cannot be accommodated in these sorts of ways.[11]

A similar lack of appreciation for how multiple distinct appraisals and re-appraisals can be involved in a single affective state undermines other alleged counterexamples. For instance, Barlassina ([2020]) discusses the case of a general and a saint, each of whom is being applauded by the populace for their achievements, and each of whom feels pride. But the general revels in (enjoys) his pride, whereas the saint is mortified—a servant of God should be humble. Here is the obvious explanation: both appraise what they did as worthy achievements, and the applause of the people as expressing admiration, and both undergo non-conceptual representations of goodness directed at themselves in light of their past actions. But the saint, alone, aware that he feels pride, appraises that fact as a betrayal of everything he stands for, and undergoes a large-magnitude representation of (non-conceptual) badness directed at himself as a result.

---

[11] Barlassina and Hayward ([2019b]) discuss a neurological case they claim to be an instance of pure valence, lacking any sort of intentional object. This was a man who had stroke damage to the primary and secondary somatosensory cortices in his right hemisphere, who was unable to feel any sensations of pain in his left hand or arm, but in whom the negative affective component of pain resulting from laser stimulation of the left hand was preserved, as discussed in (Ploner et al. [1999]). However, the quotation Barlassina and Hayward provide from this paper omits a phrase that is deeply problematic for their 'pure valence' interpretation. Here is the actual passage quoted in full, with the omitted phrases picked out in bold (only the first omission matters for our purposes): 'the patient spontaneously described a "clearly unpleasant" intensity **dependent feeling emerging from an ill-localized and extended area "somewhere between fingertip and shoulder",** that he wanted to avoid. The **fully cooperative and eloquent** patient was completely unable to further describe quality, localization and intensity of the perceived stimulus. Suggestions from a given word list containing "warm", "hot", "cold", "touch", "burning", "pinprick-like", "slight pain", "moderate pain" and "intense pain" were denied' (Ploner et al. [1999], p. 213). Far from being a case of pure undirected valence, in fact there was a vaguely specified region of the body in which something bad was felt to be happening. Cochrane ([2019]) makes a similar point.

The same failing infects Barlassina's ([2020]) examples of desires that can feel good or bad, depending on context. For an evaluativist, all felt desires will represent their objects as non-conceptually good (often in addition to representing them conceptually as good, of course). But that does not mean that all desires themselves feel good. For if at the same time one is aware that the object of desire will not be achieved, or conflicts with something else that one wants even more, then that will result in feelings of frustration (negative valence directed at the situation). When there is an expectation that the desire will be satisfied, in contrast, that fact will be appraised as good, and one will experience positive valence directed at the situation.

In further counterexamples, Barlassina and Hayward ([2019a]) are mistaken about the evaluative contents said to be represented, according to evaluativists. In fact, they conflate conceptual with non-conceptual forms of evaluative content. For example, they contrast the visual experiences of two people looking at someone's bad injury, one of whom (a trainee doctor) finds the experience unpleasant whereas the other of whom (an expert in injuries of that sort) does not; and yet, it is said, the visual experience of both represents the injury as being bad. Indeed, it does: conceptually bad. In contrast, the novice may appraise the injuries to be disgusting (issuing in a non-conceptual representation of badness) whereas the expert has become habituated to blood and no longer does.

The same problem infects Barlassina's ([2020]) example of the person who sees (looking in the mirror) that the dentist has badly damaged his mouth (seeing that his teeth are misaligned, for example), while experiencing no negative affect at the sight due to the continuing effects of the aesthetic. For again: the former is a conceptual recognition of damage, or a non-conceptual representation of something that implies damage (misalignment). No evaluativist would identify such contents with negative valence.

In addition, Barlassina ([2020]) makes a lot of play with the fact that our best theories of imaginative prospection, of the sort that is agreed to underlie valence-based decision making, involves simulated sensory experience. This is said to be problematic for evaluativism, which holds that the objects of appraisal are generally items and events in the world; and it is said to support experience-directed imperativism, instead. This argument is weak, however. Everyone can agree that prospection involves sensory images of various sorts, supported by the same brain regions that are involved in sensory experience. But an evaluativist will insist that it is the represented contents of those experiences and images that are appraised for value, not the experiences themselves. Indeed, one would think that an imperativist ought to say the same thing! This is because the appraisals that issue in positive and negative valence in real-world (non-simulated) situations are plainly responding to the represented contents of the experiences (for example, the approaching bear, in the case of fear), even if the output of the appraisal process is an experience-directed imperative 'less of me!'. When one then simulates an experience of that sort, experientially imagining being approached by a ferocious-looking bear, it should be the very same worldly content that gets appraised.

Finally, and in a more scientific vein, Barlassina and Hayward ([2019a]) also appeal to the sorts of dissociations between pleasure and world-directed motivation emphasized in the cognitive neuroscience literature (Berridge and Kringelbach [2008], [2013]; Pool et al. [2016]). An animal (or a person) can be motivated to do something, without any enjoyment and without any anticipation of enjoyment. And likewise an animal (or person) can enjoy something (for example, a sugar solution) when presented with it, while having no motivation to pursue it. These findings are said to be problematic for evaluativism because of the latter's world-directed (as opposed to experience-directed) nature. But in fact the dissociations in question provide one of the main empirical demonstrations of the distinction between direct motivations for action resulting from appraisal mechanisms (these are valence-free in the first instance) and indirect ones (which rely on positive or negative valence). This is, in fact, one of the crucial items of agreed scientific knowledge about affect mentioned in section 2, and is fully consistent with evaluativism.

6. Conclusion

This article has compared the strengths and weaknesses of two theories of the nature of affective valence. According to evaluativism, valence is a non-conceptual representation of adaptive value and disvalue. According to experience-directed imperativism, in contrast, valence has non-conceptual imperative content. We have seen that imperativism appears unnecessary—since valence is at least a representation of adaptive value—and that it confronts severe difficulties in accommodating well-established findings in cognitive neuroscience. Evaluativism, in contrast, is consistent with those findings, and faces no difficulties of its own that cannot readily be overcome.

Acknowledgements

Department of Philosophy
University of Maryland
College Park, MD, USA
pcarruth@umd.edu

References

Adolphs, R. and Anderson, D. J. [2018]: *The Neuroscience of Emotion*, Princeton, NJ: Princeton University Press.

Ashar, Y., Chang, L. J. and Wager, T. [2017]: 'Brain Mechanisms of the Placebo Effect: An Affective Appraisal Account', *Annual Review of Clinical Psychology*, **13**, pp. 73–98.

Aydede, M. [2014]: 'How to Unify Theories of Sensory Pleasure: An Adverbialist Proposal', *Review of Philosophy and Psychology*, **5**, pp. 119–33.

Bain, D. [2013]: 'What Makes Pains Unpleasant?', *Philosophical Studies*, **166**, pp. 69–89.

Bain, D. [2017]: 'Evaluativist Accounts of Pain's Unpleasantness', in J. Corns (*ed.*), *Routledge Handbook of Philosophy of Pain*, Oxford: Routledge.

Bain, D. [2019]: 'Why Take Painkillers?', *Noûs*, **53**, pp. 462–90.

Barlassina, L. [2020]: 'Beyond Good and Bad: Reflexive Imperativism, Not Evaluativism, Explains Valence', *Thought*, **9**, pp. 274–84.

Barlassina, L. and Hayward, M. [2019a]: 'Loopy Regulations: The Motivational Profile of Affective Phenomenology', *Philosophical Topics*, **47**, pp. 233–61.

Barlassina, L. and Hayward, M. [2019b]: 'More of Me! Less of Me! Reflexive Imperativism about Affective Phenomenal Character', *Mind*, **128**, pp. 1013–44.

Barrett, L. and Bar, M. [2009]: 'See It with Feeling: Affective Predictions during Object Perception', *Philosophical Transactions of the Royal Society* B, **364**, pp. 1325–34.

Barrett, L., Lindquist, K., Bliss-Moreau, E., Duncan, S., Gendron, M., Mize, J. and Brennan, L. [2007]: 'Of Mice and Men: Natural Kinds of Emotions in the Mammalian Brain?', *Perspectives on Psychological Science*, **2**, pp. 297–312.

Bartra, O., McGuire, J. and Kable, J. [2013]: 'The Valuation System: A Coordinate-Based Meta-analysis of BOLD fMRI Experiments Examining Neural Correlates of Subjective Value', *NeuroImage*, **76**, pp. 412–27.

Berridge, K. and Kringelbach, M. [2008]: 'Affective Neuroscience of Pleasure: Reward in Humans and Animals', *Psychopharmacology*, **199**, pp. 457–80.

Berridge, K. and Kringelbach, M. [2013]: 'Neuroscience of Affect: Brain Mechanisms of Pleasure and Displeasure', *Current Opinion in Neurobiology*, **23**, pp. 294–303.

Botvinick, M., Huffstetler, S. and McGuire, J. [2009]: 'Effort Discounting in Human Nucleus Accumbens', *Cognitive, Affective, and Behavioral Neuroscience*, **9**, pp. 16–27.

Boyd, R. [1991]: 'Realism, Anti-foundationalism and the Enthusiasm for Natural Kinds', *Philosophical Studies*, **61**, pp. 127–48.

Boyd, R. [1999]: 'Kinds, Complexity, and Multiple Realization', *Philosophical Studies*, **95**, pp. 67–98.

Carruthers, P. [2015]: *The Centered Mind*, Oxford: Oxford University Press.

Carruthers, P. [2018]: 'Valence and Value', *Philosophy and Phenomenological Research*, **97**, pp. 658–80.

Carruthers, P. and Williams, D. M. [2019]: 'Comparative Metacognition', *Animal Behavior and Cognition*, **6**, pp. 278–88.

Cochrane, T. [2019]: *The Emotional Mind*, Cambridge: Cambridge University Press.

Cona, G., Scarpazza, C., Sartori, G., Moscovitch, M. and Bisiacchi, P. [2015]: 'Neural Bases of Prospective Memory', *Neuroscience and Biobehavioral Reviews*, **52**, pp. 21–37.

Corbetta, M., Patel, G. and Shulman, G. [2008]: 'The Reorienting System of the Human Brain: From Environment to Theory of Mind', *Neuron*, **58**, pp. 306–24.

Cutter, B. and Tye, M. [2011]: 'Tracking Representationalism and the Painfulness of Pain', *Philosophical Issues*, **21**, pp. 90–109.

Cutter, B. and Tye, M. [2014]: 'Pain as Reasons: Why It Is Rational to Kill the Messenger', *Philosophical Quarterly*, **64**, pp. 423–33.

de Araujo, I., Rolls, E., Velazco, M., Margot, C. and Cayeux, I. [2005]: 'Cognitive Modulation of Olfactory Processing', *Neuron*, **46**, pp. 671–79.

Dehaene, S. [1997]: *The Number Sense*. London: Penguin.

Dokic, J. [2012]: 'Seeds of Self-knowledge: Noetic Feelings and Metacognition', in M. Beran, J. Brandl, J. Perner and J. Proust (*eds*), *Foundations of Metacognition*, Oxford: Oxford University Press.

Dretske, F. [1988]: *Explaining Behavior*, Cambridge, MA: MIT Press.

Ekman, P. [1992]: 'Are There Basic Emotions?', *Psychological Review*, **99**, pp. 550–53.

Eldar, E., Rutledge, R., Dolan, R. and Niv, Y. [2016]: 'Mood as Representation of Momentum', *Trends in Cognitive Sciences*, **20**, pp. 15–24.

Garcia, J., Kimeldorf, D. and Koelling, R. [1955]: 'Conditioned Aversion to Saccharin Resulting from Exposure to Gamma Radiation', *Science*, **122**, pp. 157–58.

Gilbert, D. and Wilson, T. [2005]: 'Affective Forecasting: Knowing What to Want', *Current Directions in Psychological Science*, **14**, pp. 131–34.

Gilbert, D. and Wilson, T. [2007]: 'Prospection: Experiencing the Future', *Science*, **317**, pp. 1351–54.

Grabenhorst, F. and Rolls, E. [2011]: 'Value, Pleasure, and Choice in the Ventral Prefrontal Cortex', *Trends in Cognitive Sciences*, **15**, pp. 56–67.

Greenwald, A. and De Houwer, J. [2017]: 'Unconscious Conditioning: Demonstration of Existence and Difference from Conscious Conditioning', *Journal of Experimental Psychology: General*, **146**, pp. 1705–21.

Gruber, R., Schiestl, M., Boeckle, M., Frohnwieser, A., Miller, R., Gray, R. D., Clayton, N. and Taylor, A. H. [2019]: 'New Caledonian Crows Use Mental Representations to Solve Metatool Problems', *Current Biology*, **29**, pp. 686–92.

Gruberg, E., Dudkin, E., Wang, Y., Marin, G., Salas, C., Sentis, E., Letelier, J., Mpodozis, J., Malpeli, J., Cui, H., Ma, R., Northmore, D. and Udin, S. [2006]: 'Influencing and Interpreting Visual Input: The Role of a Visual Feedback System', *Journal of Neuroscience*, **26**, pp. 10368–71.

Heathwood, C. [2007]: 'The Reduction of Sensory Pleasure to Desire', *Philosophical Studies*, **133**, pp. 23–44.

Heyes, C. [2015]: 'Animal Mindreading: What's the Problem?', *Psychonomic Bulletin and Review*, **22**, pp. 313–27.

Izard, C. [2007]: 'Basic Emotions, Natural Kinds, Emotion Schemas, and a New Paradigm', *Perspectives on Psychological Science*, **2**, pp. 260–80.

Izard, V., Sann, C., Spelke, E. and Streri, A. [2009]: 'Newborn Infants Perceive Abstract Numbers', *Proceedings of the National Academy of Sciences*, **106**, pp. 10382–85.

Jordan, K., MacLean, E. and Brannon, E. [2008]: 'Monkeys Match and Tally Quantities across Senses', *Cognition*, **108**, pp. 617–25.

Juechems, K. and Summerfield, C. [2019]: 'Where Does Value Come From?', *Trends in Cognitive Sciences*, **23**, pp. 836–50.

Kirk-Smith, M., Van Toller, C. and Dodd, G. [1983]: 'Unconscious Odor Conditioning in Human Subjects', *Biological Psychology*, **17**, pp. 221–31.

Klein, C. [2015]: *What the Body Commands*, Cambridge, MA: MIT Press.

Knudsen, E. [2011]: 'Control from Below: The Role of a Midbrain Network in Spatial Attention', *European Journal of Neuroscience*, **33**, pp. 1961–72.

Kobayashi, S., Kojima, S., Yamanaka, M., Sadamoto, H., Nakamura, H., Fujito, Y., Kawai, R., Sakakibara, M. and Ito, E. [1998]: 'Operant Conditioning and Escape Behavior in the Pond Snail, *Lymnaea stagnalis*', *Zoological Science*, **15**, pp. 683–90.

Kriegel, U. [2007]: 'The Same-Order Monitoring Theory of Consciousness', *Synthesis Philosophica*, **22**, pp. 361–84.

Krieglmeyer, R., Deutsch, R., De Houwer, J. and De Raedt, R. [2010]: 'Being Moved: Valence Activated Approach-Avoidance Behavior Independently of Evaluation and Approach-Avoidance Intentions', *Psychological Science*, **21**, pp. 607–13.

Lambie, J. and Marcel, A. [2002]: 'Consciousness and the Varieties of Emotion Experience', *Psychological Review*, **109**, pp. 219–59.

LeDoux, J. [2012]: 'Rethinking the Emotional Brain', *Neuron*, **73**, pp. 653–76.

Levy, D. and Glimcher, P. [2012]: 'The Root of All Value: A Neural Common Currency for Choice', *Current Opinion in Neurobiology*, **22**, pp. 1027–38.

Martínez, M. [2011]: 'Imperative Content and the Painfulness of Pain', *Phenomenology and the Cognitive Sciences*, **10**, pp. 67–90.

Martínez, M. [2015]: 'Disgusting Smells and Imperativism', *Journal of Consciousness Studies*, **22**, pp. 191–200.

Millikan, R. [1989]: 'Biosemantics', *Journal of Philosophy*, **86**, pp. 281–97.

Millikan, R. [1995]: 'Pushmi-Pullyu Representations', in J. E. Tomberlin (*ed*.), *AI, Connectionism and Philosophical Psychology*, Atascadero, CA: Ridgeview, pp. 185–200.

Miloyan, B. and Suddendorf, T. [2015]: 'Feelings of the Future', *Trends in Cognitive Science*, **19**, pp. 196–200.

Momennejad, I. and Haynes, J.-D. [2012]: 'Human Anterior Prefrontal Cortex Encodes the "What" and "When" of Future Intentions', *NeuroImage*, **61**, pp. 139–48.

Moors, A., Ellsworth, P., Scherer, K. and Frijda, N. [2013]: 'Appraisal Theories of Emotion: State of the Art and Future Development', *Emotion Review*, **5**, pp. 119–24.

Mysore, S. and Knudsen, E. [2013]: 'A Shared Inhibitory Circuit for Both Exogenous and Endogenous Control of Stimulus Selection', *Nature Neuroscience*, **16**, pp. 473–78.

Panksepp, J. [1998]: *Affective Neuroscience*, Oxford: Oxford University Press.

Panksepp, J. and Watt, D. [2011]: 'What Is Basic about Basic Emotions? Lasting Lessons from Affective Neuroscience', *Emotion Review*, **3**, pp. 387–96.

Penn, D., Holyoak, K. and Povinelli, D. [2008]: 'Darwin's Mistake: Explaining the Discontinuity between Human and Nonhuman Minds', *Behavioral and Brain Sciences*, **30**, pp. 109–78.

Persichetti, A., Aguirre, G. and Thompson-Schill, S. [2015]: 'Value Is in the Eye of the Beholder: Early Visual Cortex Codes Monetary Value of Objects during a Diverted Attention Task', *Journal of Cognitive Neuroscience*, **27**, pp. 893–901.

Pfabigan, D., Alexopoulos, J., Bauer, H. and Sailer, U. [2011]: 'Manipulation of Feedback Expectancy and Valence Induces Negative and Positive Reward Prediction Error Signals Manifest in Event-Related Potentials', *Psychophysiology*, **48**, pp. 656–64.

Plassmann, H., O'Doherty, J., Shiv, B. and Rangel, A. [2008]: 'Marketing Actions Can Modulate Neural Representations of Experienced Pleasantness', *Proceedings of the National Academy of Sciences USA*, **105**, pp. 1050–54.

Ploner, M., Freund, H.-J. and Schnitzler, A. [1999]: 'Pain Affect without Pain Sensation in a Patient with a Postcentral Lesion', *Pain*, **81**, pp. 211–14.

Pool, E., Sennwald, V., Delplanque, S., Brosch, T. and Sander, D. [2016]: 'Measuring Wanting and Liking from Animals to Humans: A Systematic Review', *Neuroscience and Biobehavioral Reviews*, **63**, pp. 124–42.

Proust, J. [2014]: *The Philosophy of Metacognition*, Oxford: Oxford University Press.

Rolls, E. [2014]: *Emotion and Decision Making Explained*, Oxford: Oxford University Press.

Rolls, E. [2015]: 'Limbic Systems for Emotion and for Memory, but No Single Limbic System', *Cortex*, **62**, pp. 119–57.

Rupert, R. [2018]: 'Representation and Mental Representation', *Philosophical Explorations*, **21**, pp. 204–25.

Russell, J. [2003]: 'Core Affect and the Psychological Construction of Emotion', *Psychological Review*, **110**, pp. 145–72.

Scherer, K. [2009]: 'The Dynamic Architecture of Emotion: Evidence for the Component Process Model', *Cognition and Emotion*, **23**, pp. 1307–51.

Schultz, W., Dayan, P. and Montague, P. [1997]: 'A Neural Substrate of Prediction and Reward', *Science*, **275**, pp. 1593–99.

Seligman, M., Railton, P., Baumeister, R. and Sripada, C. [2013]: 'Navigating into the Future or Driven by the Past', *Perspectives on Psychological Science*, **8**, pp. 119–41.

Shea, N. [2014]: 'Reward Prediction Error Signals Are Meta-representational', *Noûs*, **48**, pp. 314–41.

Shea, N. [2018]: *Representation in Cognitive Science*, Oxford: Oxford University Press.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T., Cohen, J. D. and Botvinick, M. [2017]: 'Toward a Rational and Mechanistic Account of Mental Effort', *Annual Reviews in Neuroscience*, **40**, pp. 99–124.

Shenhav, A., Straccia, M., Cohen, J. D. and Botvinick, M. [2014]: 'Anterior Cingulate Engagement in a Foraging Context Reflects Choice Difficulty, Not Foraging Value', *Nature Neuroscience*, **17**, pp. 1249–54.

Shenhav, A., Straccia, M., Musslick, S., Cohen, J. D. and Botvinick, M. [2018]: 'Dissociable Neural Mechanisms Track Evidence Accumulation for Selection of Attention versus Action', *Nature Communications*, **9**, pp. 1–10.

Smith, J. D., Couchman, J. and Beran, M. [2014]: 'Animal Metacognition: A Tale of Two Comparative Psychologies', *Journal of Comparative Psychology*, **128**, pp. 115–31.

Taylor, A., Elliffe, D., Hunt, G. and Gray, R. [2010]: 'Complex Cognition and Behavioral Innovation in New Caledonian Crows', *Proceedings of the Royal Society* B, **277**, pp. 2637–43.

van Dantzig, S., Pecher, D. and Zwaan, R. [2008]: 'Approach and Avoidance as Action Effects', *Quarterly Journal of Experimental Psychology*, **61**, pp. 1298–306.

von Bayern, A., Danel, S., Auersperg, A., Mioduszewska, B. and Kacelnik, A. [2018]: 'Compound Tool Construction by New Caledonian Crows', *Nature Scientific Reports*, **8**, pp. 1–8.

Wallis, J. [2007]: 'Orbito-Frontal Cortex and Its Contribution to Decision-Making', *Annual Review of Neuroscience*, **30**, pp. 31–56.

Winstanley, C. and Floresco, S. [2016]: 'Deciphering Decision Making: Variation in Animal Models of Effort- and Uncertainty-Based Choice Reveals Distinct Neural Circuitries Underlying Core Cognitive Processes', *Journal of Neuroscience*, **36**, pp. 12069–79.

Yarkoni, T., Poldrack, R., Nichols, T., Van Essen, D. and Wager, T. [2011]: 'Large-Scale Automated Synthesis of Human Functional Neuroimaging Data', *Nature Methods*, **8**, pp. 665–70.