Special Issue: Consciousness science and its theories
A possible evolutionary function of phenomenal conscious experience of pain

Oren Kolodny[1], Roy Moyal[2], and Shimon Edelman ⓘ [2,*]

[1]Department of Ecology, Evolution, and Behavior, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Edmond J. Safra Campus - Givat Ram, Jerusalem 9190401, Israel; [2]Department of Psychology, Uris Hall, Ithaca, NY 14853, USA

*Correspondence address. Department of Psychology, Uris Hall, Ithaca, NY 14853, USA. E-mail: edelman@cornell.edu

Abstract

Evolutionary accounts of feelings, and in particular of negative affect and of pain, assume that creatures that feel and care about the outcomes of their behavior outperform those that do not in terms of their evolutionary fitness. Such accounts, however, can only work if feelings can be shown to contribute to fitness-influencing outcomes. Simply assuming that a learner that feels and cares about outcomes is more strongly motivated than one that does is not enough, if only because motivation can be tied directly to outcomes by incorporating an appropriate reward function, without leaving any apparent role to feelings (as it is done in state-of-the-art engineered systems based on reinforcement learning). Here, we propose a possible mechanism whereby pain contributes to fitness: an actor-critic functional architecture for reinforcement learning, in which pain reflects the costs imposed on actors in their bidding for control, so as to promote honest signaling and ultimately help the system optimize learning and future behavior.

Keywords: theories and models; pain; evolution; honest signaling

> But should you fail to keep your kingdom
> And, like your father before you, come
> Where thought accuses and feeling mocks,
> Believe your pain.
>
> *Alonso to Ferdinand*
> — W. H. Auden

Introduction

Why does pain hurt? Why do some things feel good and others bad? And why is it that being conscious always feels like something? An obvious place to look for answers to these questions is evolution—assuming that consciousness makes a difference in the functioning of creatures that are capable of it, as opposed to being purely epiphenomenal. This assumption is sometimes flagged as such in discussions of evolution (as in Panksepp 2005, p.35: "I assume that the evolution of consciousness was based on the ability of neural tissues to encode biological values") and of behavior (as in Frith and Metzinger 2016, p.197: "The starting assumption is that consciousness (subjective experience), rather than being an epiphenomenon, has a causal role in the optimization of certain human behaviors").

In this article, we attempt to substantiate one aspect of this assumption, by advancing a hypothesis concerning the evolutionary function of pain, which has been inspired by the Handicap Principle (Zahavi 1975; Zahavi and Zahavi 1997) and by the more recent developments in the theory of honest

signaling (Higham 2014; Zahavi 2018). Specifically, our hypothesis offers an account for the function of pain in learning and behavioral regulation and identifies a mechanism in the vertebrate nervous system that may implement that function.

What, then, is pain good for? The medical consensus considers pain as a kind of alarm mechanism that signals damage to the organism's tissue integrity (Walters and Williams 2019) and general homeostasis (Craig 2005; Damasio and Carvalho 2013; if pain indeed indicates any deviation from homeostasis, it may be why "mental" (including social) and physical pain have the same brain basis and physiology; Craig 2002; Eisenberger 2012; Durso *et al.* 2015). Nesse and Schulkin (2019, p.1) open their review of the evolutionary medicine perspective on pain by noting that it "always seems like a problem, but usually, it is part of the solution." They then proceed to explain the occasional seeming excessiveness of pain by invoking the "smoke detector principle," according to which raising an occasional false alarm is better than missing the fire. Given that knowing that one's home is on fire is without doubt more conducive to one's survival and eventual total fitness than not knowing it, Nesse and Schulkin (2019) conclude that pain has a clear evolutionary basis. Raising the stakes, Metzinger (2017 p.252) postulates in passing a broader evolutionary role for suffering (a relative of pain; Fink 2011, p.46) than merely serving as a distress signal: "functionally speaking, suffering is necessary for autonomous self-motivation and the emergence of truly intelligent behavior."

Such appeals to evolution can, however, only be effective if it is demonstrated that *feelings* such as fear—over and above merely yoking the smoke alarm directly to an evacuation order—indeed contribute to fitness-influencing outcomes. Simply assuming that an agent that feels and cares about outcomes is more strongly motivated than one that does not is not enough. Consider an artificial autonomous system, such as a self-driving car, whose reinforcement learning reward function is made to value danger avoidance, and in which motivation directly drives action. Such a system would seem to be able to do well without leaving any apparent role to feelings in its functioning. This would make it into a kind of philosophical zombie (Kirk 2019)—namely, an AI zombie. More generally, it has been recently argued that functionally effective AI systems may not have to (and, normatively speaking, should not) be made to incorporate suffering (Agarwal and Edelman 2020).

Somewhat surprisingly, the possibility of effective functioning without the involvement of pain in the regulation of behavior does not seem to have been raised often, let alone decided upon, in the literature (i.e., outside the vast ongoing debate about whether or not zombies, which seem to be philosophically conceivable, are therefore also physically possible Worley 2003; Márton 2019; Polcyn 2020; it is not our intention in this article to weigh in on that debate). One notable exception is Arbilly and Lotem (2017), who write (p.2): "[…] we should first ask what it means to 'feel' and how a mechanism that 'feels' hunger may be adaptive for animals. This question is important because in theory, a hunger system can also work by conditional rules or stimulus-response switches without involving the 'feeling' of hunger." In the remainder of their paper, however, Arbilly and Lotem (2017) pivot away from the question of the utility of *feeling* hunger to the utility of representing it in memory.

In a similar vein, Carruthers (2018, p.661) states that "Two issues are worth brief discussion. One is whether goals and intentions can motivate action independently of affective states. The other is whether cognitive states alone (e.g., beliefs

about what is good or bad) can motivate action." Carruthers answers both these questions in the negative, relying in each case on arguments that invoke human-specific brain structures and processes. The question therefore still stands: if it is at all possible to function well enough without feeling a thing, like an AI zombie, why are we not like that?

Phenomenality, Feelings, and Pain

Given our just-restated question, the first order of business should be to establish that its premise holds: that to be phenomenally aware means always to feel and care (and occasionally experience pain), and not just to perceive and act (and merely reflexively withdraw when pricking a finger on a thorn). To that end, in this section, we state the place of phenomenal awareness among the various types of consciousness and mention the evidence suggesting that feeling and caring are part of the package, as phenomenal awareness in a typical human has obligatory aspects of affect and valence.

Briefly, phenomenal awareness is a basic type of consciousness, the capacity for which we share with other animals (e.g., Panksepp 2005; Edelman *et al.* 2016). Examples include the awareness of a touch, of redness, of saltiness, of breathlessness, of effort, etc. Notably, basic phenomenal awareness does *not* need to involve the complete experience of being a self, which for vertebrates like ourselves is thought to include, in addition to it, (i) a sense of situatedness in space, (ii) a sense of ownership of the body, (iii) a sense of agency over actions, and, fundamentally, (iv) a sense of being [Merker 2007; Metzinger 2003; see (Edelman 2008, 9.3.1) for a concise formulation of the components of vertebrate conscious experience and (Blanke and Metzinger 2009) for a discussion of minimal phenomenal selfhood]. Basic phenomenal awareness is thus the most stripped-down variety of consciousness that still allows for any sensorimotor experience; what may be left if that too is taken away is minimal phenomenal experience (Metzinger 2018, 2020), as in lucid dreamless sleep (Windt 2015).

Phenomenal awareness has been the subject of a number of computational theories, which attempt to explain it, for instance, in terms of probabilities of the system's state transitions (Tononi 2008; Oizumi *et al.* 2014) or the geometry of its state-space trajectories (Fekete and Edelman 2011; Moyal *et al.* 2020). These, however, are secondary to our present effort, which focuses on trying to understand the functional—specifically, evolutionary—reason for the existence of the special variety of phenomenal experience that consists of the feeling of pain. Thus, of the many possible aspects of the question "Why does pain hurt?", our main concern in this article is with one that has to do with the evolutionary reasons behind pain's hurtfulness.

That feelings such as pain are an integral part of any conscious experience is indicated by extensive evidence from psychology, neurobiology, and evolutionary science (e.g., Panksepp 2001, 2005; Lindquist *et al.* 2012; Damasio and Carvalho 2013; Feldman Barrett 2013). Feelings are, therefore, integral also to all other types of consciousness that are built on top of basic phenomenal awareness. Of course, given the extent and the scope of the brain's unconscious activity (Kihlstrom 1995), not everything that happens to us enters our awareness, not by far [as (Panksepp 2005, p.38) puts it, "felt experience stews within an enormous unconscious neural caldron"]; but everything that does is not merely informationally registered but also felt. Pain, in particular, has both informational (sensory-discriminative) and felt (affective-motivational) aspects (Auvray *et al.* 2010).

The normally obligatory integration of sense and affect frays under certain abnormal conditions. For instance, subjects presenting with pain asymbolia (Berthier *et al.* 1988; see Klein 2015; Gerrans 2020 for recent discussions) report registering the pain but not "minding" it. This condition can be brought about by a surgical alteration of brain connectivity (Carruthers 2018; Kozuch 2020). Here is how Kozuch (2020, p.695) describes pain asymbolia resulting from a surgical intervention: "That pain has both sensory and affective components is something vividly demonstrated in the effects of a cingulotomy (removal of the anterior cingulate cortex), an operation performed on patients with chronic, excruciating pain. After the operation, patients say that they still feel the pain (i.e., they have the sensory component), but that they do not mind it (i.e., they lack the affective component)[.]" Note that "feeling" in this passage refers to the informational aspect of pain, as Kozuch clarifies parenthetically.

As Nesse and Schulkin (2019) make clear, naturally occurring reduced sensitivity to pain, as in the various syndromes of innate or acquired pain deficiency, leads to an extremely negative prognosis for the individual's viability and longevity. If anything, this strengthens the case for the involvement of phenomenal pain (over and above information) in maintaining evolutionary fitness, which is where we turn next.

An Evolutionary-Functional Take on Pain

Our concern here is not with the evolutionary history of pain as such (about which "remarkably little is known"; Walters and Williams 2019, p.1) but rather with the evolutionary function of pain (and affect in general) as an aspect of phenomenal awareness. Little work has been devoted to this question to date, most of it focusing on the distinction between representational and motivational accounts (for a learning-based functional account of consciousness that is orthogonal both to this distinction and to the hypothesis that we develop here, see Cleeremans *et al.* 2020). According to the former, conscious states are uniquely useful representations of behaviorally important information (Feinberg and Mallatt 2013), or even evolutionary fitness (Kozuch 2020). According to the latter, conscious states are uniquely useful because of their motivational power (Panksepp 2005; Anderson 2019).

Frith and Metzinger (2016, p.197) offer an intriguing twist on the motivational approach by positing an evolutionary role for the experience of regret: "[…] the recognition that one could and should have acted otherwise). Regret is a powerful, negative emotion that is suggested to integrate group norms and preferences with those of the individual. The transparent and embodied nature of the experience of regret ensures that cultural norms become an inescapable part of the self-narrative. The conclusion is that conscious experience is necessary for optimizing flexible intrapersonal interactions and for the emergence of cumulative culture." This may be so, but why does one have to experience regret as pain, instead of simply making a mental note not to repeat the act that led to it?

Levels of understanding

A complete and integrated understanding of the nature of pain (let alone of affective experience in general) would have to address this "why" question alongside many others, spanning multiple aspects of the problem and several levels of explanation. The classical level distinctions that apply are: proximate vs. ultimate causes (Mayr 1961; Laland *et al.* 2011); explanations

of causation, survival value, ontogeny, and evolution (Tinbergen 1963; Bateson and Laland 2013); and the levels of computation, representation and algorithm, and implementation (Marr and Poggio 1977; Edelman 2012).

Focusing on evolutionary questions arising in connection with pain can help us understand how existing computational- and algorithmic-level accounts can be built upon to arrive at a more complete explanation. A particularly promising family of computational theories of affect is based on the predictive processing (PP) framework, which considers the brain to be a dynamical, hierarchical, Bayesian hypothesis-testing mechanism, whose overarching goal is prediction error minimization (PEM; see Fernandez Velasco and Loev 2020, Section 2) and (Hohwy 2020) for recent in-depth reviews).

Theories of affect based on the PP paradigm often focus on valence and may refer to pain only in passing, treating it as a synonym for negative valence. For instance, Van de Cruys (2017) equates valence with the first derivative of prediction error over time, with positive valence corresponding to a reduction of prediction errors, which may stem from the agent's own actions; Joffily and Coricelli (2013) offer a similar account rooted in the free-energy principle advanced by Friston (2010). A recent synthesis of several PP approaches, the Affective Inference Theory of Fernandez Velasco and Loev (2020), holds that valence corresponds to the expected rate of prediction error reduction.

While Fernandez Velasco and Loev (2020) include a comprehensive and detailed discussion of pain as reflected upon by their theory, they literally take for granted the evolutionary origins of affect in general and of pain in particular: "If we grant that evolutionary pressure has made sure that allostasis and PEM are two sides of the same coin […], then valence (Rate) can be used to maintain the policies that minimize Error over time" (Fernandez Velasco and Loev 2020, p.20). This leaves open Tinbergen's questions concerning the adaptive value of pain and the process whereby it came to be that way. It is these questions that motivate us in the present effort.

To be fully useful, answers to evolutionary questions must engage with Marr's levels (which are themselves interdependent; Edelman 2012). Thus, while Nesse and Schulkin (2019) explicitly mention Tinbergen's four questions, their casting of pain as a metaphorical smoke detector offers no computational-level insights into what we consider to be the central question about pain: why is it not merely informational?

Useful hints about evolution can sometimes be gleaned from an implementation-level analysis of a system. Thus, Nesse and Schulkin (2019, p.2) remark that "no single mechanism regulates pain sensitivity in general, and it further suggests that different kinds of nociception diverged long ago or that they had separate origins." For us, this implies that a unifying evolutionary reason, if any, for the experiential quality of pain, must be sought on a more abstract level than that of neural mechanism (although a valid functional explanation would have to be compatible with known anatomy and physiology). Thus, the explanation we seek is functional–computational, having to do specifically with signaling (Zahavi and Zahavi 2012) among subpersonal computational processes involved in individual decision-making and learning.

The conceptual framework: honest signaling

As we already noted, the standard account of pain, which holds it to be a signal sent by a damage-monitoring process to the central controller, fails to explain why pain is hurtful and not just informational (here, as earlier, we invoke the distinction

between sensory and affective aspects of pain; Auvray *et al.* 2010). But what if pain is not a signal as such, but rather the manifestation of a cost, imposed on a signaling process to ensure its honesty?

In evolutionary theory, the conceptual framework of costly signaling (Higham 2014; Laidre and Johnstone 2013) grew out of the Handicap Principle (Zahavi 1975; Grafen 1990; Zahavi and Zahavi 1997). A classical example, discussed by Darwin in the context of sexual selection, is the peacock's tail. As a clear handicap to its owner, it presents a paradox: how and why could such a trait emerge and remain stable? The intuitive explanation offered by Zahavi (1975) holds that the handicap serves as a hard-to-fake cost, which only high-quality individuals can afford, and which therefore helps the intended recipient of the signal, the peahen, assess its reliability, ultimately boosting the peacock's fitness.

Another extensively studied case is that of nestling begging (e.g., Godfray 1991; Rodríguez-Gironés *et al.* 1996; Godfray and Johnstone 2000). The ruckus raised by chicks in a nest, which compete for the attention of a parent that has just arrived with a morsel of food, is energetically costly. It should be noted that, unlike the peacock tail, this is not a case of differential quality, but rather of differential need (which in the context of reinforcement learning is the same as differential value for each of the players of the expected reward if their signal is accepted and acted upon). The presence of a cost raises the question of the joint dynamics of the chicks' begging (which, also unlike the growing of a tail by a peacock, can be regulated in real time) and the parent's response (which can be regulated as well). Rodríguez-Gironés *et al.* (1996, p.14637) discuss the emergence and the evolutionary stability of the begging-and-response dynamics, remarking that "Since nonsignaling is presumably the ancestral condition, this raises the question of what led to the evolutionary change toward the present signaling equilibrium"; the account they offer shows how stable cost-regulated honest signaling could have evolved in this situation.

It is important to realize that signaling can be carried out not only between but also within individuals. The concepts of signaling and honesty or reliability are thus directly applicable to the questions of the evolutionary origins and role of pain. Harris *et al.* (2014) invoked the idea of intra-individual costly signaling as an explanation for the otherwise puzzling neurotoxicity of neurotransmitters in the central nervous system: "In the CNS [central nervous system], minor changes in the concentration of neurotransmitters such as glutamate or dopamine can lead to neurodegenerative diseases. We present an evolutionary perspective on the function of neurotransmitter toxicity in the CNS. We hypothesize that neurotransmitters are selected because of their toxicity, which serves as a test of neuron quality and facilitates the selection of neuronal pathways."

A key observation here is that the function of the toxicity of a neurotransmitter is logically independent of its efficacy and cannot therefore be subsumed into the dynamics of the latter, i.e., of synaptic modification for purposes of learning. This insight can be applied to the case of phenomenally experienced pain. Consider the withdrawal reflex. Disengagement from a noxious stimulus is enacted immediately and does not, in principle, require that the stimulus also be experienced as painful. Thus, the role of pain here cannot be merely to motivate an immediate withdrawal, which is in fact taken care of by the reflex arc. Neither can the role of pain be merely to instill later aversion to the stimulus: aversion, like motivation, is informational and does not have to involve phenomenal experience.

Could pain play a role in reinforcement learning? By definition, learning from behavioral outcomes requires that the blame or credit for the results of an action be apportioned to its causal predecessors, giving rise to the so-called Credit Assignment Problem that is central to AI [first formulated by Minsky (1961, p.20); see Schmidhuber (2015) for an exhaustive review]. Clearly, credit assignment is an informational function, and so, just like motivation and aversion, it does not need phenomenal pain to mediate it.

In this sense, it is not surprising that phenomenal pain is doubly dissociable from mere information about physiological distress [as indicated by the condition of pain asymbolia (Klein 2015; Gerrans 2020) and what has been described as its opposite (Ploner *et al.* 1999); Hagiwara *et al.* (2020) report a curious condition that seems to be related to both]. The crucial observation is that if pain were entirely informational, there would be no special contribution for it to make to the control of behavior, as its function would be entirely reducible to its projections onto the relevant non-affective dimensions (motivation, negative tropism, credit, etc.).

A general conclusion from these considerations is that the special function of pain can be neither in agreement nor at odds with motivation, tropism, or credit, which are all informational. Consequently, the special contribution of pain *qua* feeling must be *orthogonal* to motivation and to credit—just as the toxicity of neurotransmitters is functionally orthogonal to their efficacy in conveying information across the synaptic cleft.

We are now in a position to introduce our hypothesis: the role of the phenomenal experience of pain is managing certain functional components of the behavioral control system—namely, processes each of which is associated with honing and promoting a particular context-specific action plan. What could the dimension on which such *actor* processes are differentiated be about? As we just concluded, it cannot be defined exclusively in terms of outcomes (as this would make it functionally indistinguishable from reward). We therefore propose that this independent dimension has to do with the actor's *confidence* in the plan that it embodies.

Consider a situation in which you have spied some blackberries deep in a bramble thicket and are attempting to reach for them. The present pain caused by the thorns and the memories of past pain inflicted by thumbtacks and nails motivate you to withdraw; yet, something inside you tells you to persevere, and so you push through until the sweet objective is attained. It is now time to assess your behavior. The assessment is not just of the reward (how sweet the blackberries turned out to be) but of its relationship to your original confidence and to whatever damage you incurred in acting on it. Next time, your confidence in deciding to persevere in the face of pain will be higher, or lower, depending on the outcome of the assessment.

Similarly, consider the pain that you experience while holding what is proving to be a very hot cookpot. The motivation to immediately drop the pot is being weighed against the consequences of doing so. In the end, your confidence that those consequences would be much worse than getting a burn on your palm has the upper hand and you return the pot safely to the stove.

We note that the outcome of the post-incident assessment is distinct from the reward prediction error in reinforcement learning (e.g., Gershman 2017), because confidence in the sense proposed here has to do not just with the expected reward but also with a range of additional considerations, including what it takes to achieve it. This is why this notion of confidence is also distinct from prediction confidence in a generative model used

in active-inference control (a predictive coding approach derived from the free-energy principle; Kanai *et al.* 2015), where "the estimated precision (inverse variance) of the inferred policy … corresponds to the agent's confidence that the policy it is currently following is optimal" (Gershman and Uchida 2019, p.711).

The scenario that we envisage can be described in the abstract as follows. In a behavioral control system, multiple actors compete for the privilege of executing their respective policies. The competition takes the form of bidding: the price of participation is raised, such that only the better-off actors can afford it. Each actor has an interest to be chosen if it is confident in the applicability of its policy and in its ability to execute it; otherwise, it has an interest to avoid being chosen. This is because an actor can be either paid or fined after its participation, depending on the outcome. The system as a whole has an interest in choosing actors whose confidence is justified. An actor's bid to meet the cost of participation is thus a signal in the classical sense that both sides stand to gain by using it (Laidre and Johnstone 2013)—but only if the signal is honest. The honesty of signaling in this scenario is underwritten by the bidding process: actors can only gain access to output control by paying up front a cost, which is imposed uniformly on all the actors, and which the overconfident ones can, in the long run, ill-afford. This global cost is experienced as pain.

The computational mechanism: an actor-critic architecture for learning

How can the signaling scenario that we just outlined be implemented? At a minimum, in a behavioral regulation and learning system, there is an assessor, which tracks relevant data (e.g., deviations from homeostasis, tissue damage, or social–behavioral outcomes) and communicates its findings to the controller. As the assessor and the controller, being parts of the same organism, share a common fate, it may seem that the honesty of their communication is not an issue. Zahavi (2018, p.4) notes that this is not the case:

> Reliability in signaling (and hence investment in handicaps) is needed not only in order to avoid cheating in the case of a conflict of interests between signaler and receiver—as assumed by most modelers—but also in order to avoid errors in communication between individuals whose interests are identical. Whenever information is received via signals rather than from direct observation, it is important to test its reliability. Handicaps imposed by signals within the organism provide additional support for the claim that all signals need to be tested by handicaps.

Be that as it may, the "controller" in the present case resolves on a closer look into subcomponents that are in fact in competition against each other (if not against the organism as a whole)—for instance, the representations of different courses of action in a given situation. Thus, the reliability of signaling in this system does need to be ensured.

The actor-critic architecture (Barto 2013, Figure 2) that is the computational approach of choice to intrinsically motivated reinforcement learning (Baldassarre and Mirolli 2013) can help us ground this idea in a concrete model. On the classical account, the critic is charged with integrating the data from multiple sources and casting the assessment into a scalar "common currency" format (Levy and Glimcher 2012). The resulting credit/blame is then apportioned to the multiple actors, representing various state-specific behavioral choices or task sets (Koechlin 2016).

A Possible brain basis for the actor-critic architecture

In vertebrates, this abstract functional architecture maps rather closely onto the system of multiple parallel circuits that connect the basal ganglia—an evolutionarily old midbrain structure (Grillner and Robertson 2016)—via the thalamus, to the forebrain (in mammals, the cortex) and back (Alexander *et al.* 1986; Redgrave *et al.* 2013; Friend and Kravitz 2014). The wiring of this system in the mouse has recently been described in great detail by Foster *et al.* (2020), who also provided for the first time direct evidence for the closed loop-like structure of the parallel circuits. Several key properties of this system that make it relevant to our hypothesis are briefly reviewed below.

Multiple actors

The cortico-basal-ganglia-thalamocortical recurrent loop system is "one of the most important structural motifs in the mammalian brain, with nearly half of the brain participating in its elements" (Foster *et al.* 2020). It is instrumental in task-related memory maintenance, decision-making, and behavioral control, the latter functions forming a hierarchy of sensory-motor, contextual, and schemata levels (Badre and Nee 2018).

We hypothesize that the loops comprising this system implement our model's actor mechanisms, whose corresponding actions can be more or less abstract, depending on the actor's place in the control hierarchy. The number of distinct actions can be as large as the number of anatomical "domains" threaded by the parallel loops; larger, if the functional domains are graded and distributed. In the mouse, the number of anatomically defined domains is about 50:14 connected through the substantia nigra pars reticulata (SNr) and 36 through the globus pallidus external (GPe) (Foster *et al.* 2020, p.6).

Action selection

The basic mechanism for action selection in the loop system operates on the competition principle, as required by our model. Indeed Foster *et al.* (2020, p.5) attribute to the loop system the so-called executive function, which epitomizes cognitive control. Of the many possible actions, one is enabled and the others are actively inhibited, both by dedicated inhibitory pathways and by other cross-talk interconnections between the loops along their successive stations (Joel and Weiner 1994). A computational theory of this process, including the contributions of the distinct Go and NoGo pathways, has been summarized by O'Reilly (2006); a more complex functional picture is painted by Redgrave *et al.* (2010).

Learning and flexibility

Action representations embodied in the multiple parallel loops are subject to learning, as indicated by the presence of synaptic plasticity throughout this system (Wickens 2009). Action selection generally happens under the control of reward signals (Humphries and Prescott 2010; Redgrave *et al.* 2011). As discussed by Foster *et al.* (2020), in stimulus-response or habit learning, a stimulus can trigger a hitherto reliably rewarded response, even when reward value is degraded (a runaway process of reward-driven learning can lead not only to habit formation, but also to addiction, both orchestrated by basal ganglia mechanisms; Graybiel 2008; Redgrave *et al.* 2010). In contrast, action-outcome learning is goal-directed and results in behavior that is chosen on the basis of the perceived motivational value and estimated obtainability of the outcome—characteristics that parallel the process of actors bidding on access to control, as posited by our model.

Action selection in the loop system is also open-ended in the sense that novel (i.e., unpredicted) actions can be discovered and harnessed through intrinsically motivated learning (Redgrave *et al.* 2013). An explicit connection between such open-ended learning and consciousness has been proposed by Cotterill (2001, p.23), who described it as "the acquisition of new context-specific reflexes," a process in which consciousness can, but does not have to, become involved, and to which it may contribute flexibility—including, as we conjecture in the present paper, the capacity to commit to and pay now for a potential action, to be undertaken right away, or in the future, if circumstances for it arise.

Pain

Although the role of the basal ganglia in the control of aversive behavior is still poorly understood (Lloyd and Dayan 2016, p.1), their involvement in nociception and pain has been known for a while. According to Chudler and Dong (1995), basal ganglia may be involved in the "(i) sensory-discriminative dimension of pain, (ii) affective dimension of pain, (iii) cognitive dimension of pain, (iv) modulation of nociceptive information, and (v) sensory gating of nociceptive information to higher motor areas." More recent reviews attribute to the basal ganglia "a key role" in pain and analgesia (Borsook *et al.* 2010) and document their contribution to emotion in general (Pierce and Péron 2020). The breadth of this involvement is suggested by the hypothesized basal ganglia dysfunction in complex regional pain syndrome (Azqueta-Gavaldon *et al.* 2017) and the reported reorganization of the functional connectivity of basal ganglia (along with many other regions) in animal models and in patients suffering from chronic pain (Kuner and Flor 2017).

Among the better understood aspects of the basal ganglia's involvement in phenomenal pain is the role of their interaction with the posterior insula, the "primary integrative hub of interoceptive afferents," and the anterior insula, which "sits at the apex of the so-called "salience system," the neural hierarchy that signals whether and how information matters to the organism" (Gerrans 2020, p.5). A malfunction of the insula—basal ganglia circuit can result in pain asymbolia, mentioned in The conceptual framework: honest signalling section (Berthier *et al.* 1988). Interestingly, artificial stimulation of the limbic domains of basal ganglia and anterior insula can lead to a disturbance of approach-avoidance behaviors in non-human primates (Saga *et al.* 2019)—a relevant finding, if the behavioral change in this study was due to an induced pain-like affect. Finally, Fazeli and Büchel (2018) report that pain-related expectation and prediction error signals in the anterior insula are not related to aversiveness as such—a finding that seems relevant to our hypothesis of the orthogonality of affective and informational–motivational factors in action selection.

Bringing it together: the proposed explanation of the evolutionary function of pain

We can now sharpen the focus on our questions of interest and offer tentative answers. We *identify* pain with the paying of a cost that facilitate a reliable (in the sense of honest signaling theory) selection of courses of action in the face of an adverse situation. From the methodological standpoint, this claim of identity is akin to the one made by Van de Cruys (2017, p.8) in his explanation of valence: "Importantly, emotional valence is not something added to these error dynamics, it is those dynamics" (a related explanation of valence within the framework of predictive coding, offered by Fernandez Velasco and Loev (2020, p.30), likewise amounts to a claim of identity: "valence lies in expected error dynamics").

From the discussion in the preceding sections, it appears that pain is an "extra," added on top of nociceptive information that is being exchanged in the body. Tellingly, although pain's quality and intensity can be modulated by other processes, it cannot be entirely abolished, except by artificial (e.g., pharmacological or surgical) means, and its dysfunction in cases such as pain asymbolia results in impaired behavioral outcomes. This suggests that pain—i.e., the paying of the cost—plays too important an evolutionary role to be ignored by its subject at will.

1. *What is the evolutionary reason because of which pain is phenomenally felt and not merely informational?*
   The evolutionary function of pain consists in providing an additional, independently effective basis for action selection that is orthogonal to the outcome-based reward that drives reinforcement learning.
2. *What is the computational mechanism whereby pain fulfills its function?*

   The mechanism is a variant of an actor-critic architecture, in which the cost of bidding on an action differentially affects the confidence account balance of the participating actors. The "rich" (higher-balance) actors are better positioned to meet this up-front cost, which buys them preferential access to behavioral control.

Importantly, the interests of the different actors are not perfectly aligned—that, indeed, is the point of having multiple actors in the first place (cf. the theory of neuronal group selection; Edelman and Finkel 1984). Each actor maintains a confidence-related account, whose balance is used to meet the cost and is distinct from (orthogonal to) the cumulative reward that drives reinforcement learning; the functional role of positive-affect states would then be to top up the balance of the responsible actor(s). The cost of pain thus induces an implicit ranking over the actors: while the cost is the same for all actors (unlike in the case of chick begging or neuron signaling), not all actors can equally afford it. The ranking is made explicit during the bidding process: rich (high-ranked) actors are more likely to act (now and in a similar context in the future), compared to poor (low-ranked) ones. Furthermore, their account balance, along with possible selective down-modulation of the effects of certain cost "hits," favor the rich actors in pursuing longer-term goals. This contributes to the system's sophistication in learning from experience.

Some critical questions and possible answers

Some of the more obvious questions arising in connection with our hypothesis are briefly addressed below.

But why is paying the cost *felt*? Can't the system pay the cost while remaining emotionally indifferent?

While question of the *nature* of affect, and of phenomenality in general, is undoubtedly central to consciousness research, the focus of the present article is on the *function* of pain. Because arguing for the inseparability of the two is not our goal here, we will merely note that this inseparability is a central tenet of several theories of phenomenal consciousness and affect (e.g., Oizumi *et al.* 2014; Fernandez Velasco and Loev 2020), including our explanatory framework of choice, the Dynamical Emergence Theory (Moyal *et al.* 2020). According to the latter, all phenomenality, including pain and other affective processes, comes down to the dynamics of the underlying computation—

in the case of pain, of paying the cost and maintaining confidence account balance. This is because the Dynamical Emergence Theory *equates* phenomenality with certain intrinsic topological properties of the system's trajectories through its state space.

Most importantly, we know that whatever the dynamics of pain turns out to be, it must be intrinsic: it is not enough for the system's states to be interpreted externally as meeting the stipulated conditions. The absolute need for any account of consciousness in general to be intrinsic to the system in question has been pointed out by Tononi (2008, p.220) and Fekete and Edelman (2011, p.808). Interestingly, a computational distinction exists between intrinsic and external views of a system, as per Friston *et al.* (2020): "the information geometry induced in any system—whose internal states can be distinguished from external states—must acquire a dual aspect. This dual aspect concerns the (intrinsic) information geometry of the probabilistic evolution of internal states and a separate (extrinsic) information geometry of probabilistic beliefs about external states that are parameterized by internal states." This distinction expresses dual-aspect monism (Atmanspacher 2012, 2017)—a philosophical take on phenomenal experience, which we find appealing and which, according to Panksepp (2005, p.30), "saves us from various conceptual conundrums."

We note that the "Why is it felt?" question applies equally to the computational accounts of affect that are rooted in predictive processing, discussed briefly in Levels of understanding section; for those accounts too, there is ultimately no resort except equating phenomenality with whatever it is being reduced to (see the quote from Van de Cruys (2017) in the first paragraph of Bringing it together: the proposed explanation of the evolutionary function of pain section).

Why should having the experience of pain be more effective than simply hardwiring the critic's output to the suppression of the culprit actor?

It isn't necessarily. The ancestral condition in this regard might have been behavior regulation without phenomenal pain (akin to purely information-driven reinforcement learning). Once evolution has stumbled upon the kind of dynamics that amounts to the experience of pain, a new equilibrium became established (complete with the exaptation of phenomenal hurting to include all tissue damage). Rodríguez-Gironés *et al.* (1996, p.14640) sketch an analogous scenario in their discussion of the evolution of nestling begging: "[I]f, due to historical factors a species is trapped in a paradoxical ESS [evolutionarily stable state], it will normally not be able to jump to the common-sense ESS. [...] Food solicitations may have evolved in species rearing several offspring simultaneously, as a result of physical competition between siblings for the resources provided by the nondiscriminating parents. [...] [I]n multichick broods, so long as parents cannot avoid sibling competition, there is no nonsignaling Nash equilibrium." By analogy, it may be the case that in a multiple-actor setting, so long as the system cannot avoid actor competition, there is no non-pain Nash equilibrium.

Furthermore, the recourse to pain, as per the honest signaling explanation that we offer, may allow the system to be more flexible in its choice of behaviors; in particular, to trade off the tactical short-term inconvenience of experiencing pain for the advantage of attaining a strategic longer-term goal. On the one hand, endogenous modulation of pain is possible and may be evolutionarily beneficial, as hypothesized by Fink (2011, p.62), who posits an "evolutionary advantage that may be gained from suppressing the unpleasant emotional effect of pain: an animal need not act upon a pain perception or can act with a diverse range of responses." Indeed, Walters and Williams (2019, p.3) note in their review of the evolution of pain that' a complex pattern of endocannabinoid modulation of nociceptive responses has been conserved for over half a billion years, and/or [...] is a product of convergent evolution. ' On the other hand, however, the functional advantages of the capacity for pain make an evolutionary jump to its permanent abolition unlikely.

How does the pain-as-cost hypothesis relate to the basic phenomenological characteristics of pain?

Because it aims to explain the affective qualities of pain, our hypothesis is neutral with regard to its sensory aspects. On the affective side, pain's intensity corresponds to the cost that is being exacted. The location of pain in the body, which seems to be more of a sensory quality, may have to do with the sensory afferents of the actors involved (the somatotopy of the corticostriatal projections (e.g., Foster *et al.* 2020) is relevant in this regard). Working out an account of pain's secondary qualities (e.g., throbbing vs. burning) is, however, outside the scope of the present paper.

Some predictions and future work

The key prediction of our hypothesis is that actor processes, presumably implemented by the parallel loops in the basal ganglia/thalamus/cortex circuit, carry representations of their bidding strength or confidence reserves (which are distinct from the representations of reward and reward prediction error that are central to reinforcement learning). These, moreover, are depleted by experienced pain (and replenished by outcome-driven positive affect). It should be possible to isolate and study such representations, using the same techniques that have yielded evidence of prediction error and confidence being represented in the basal ganglia dopamine system (e.g., Gershman and Uchida 2019; Chen and Goldberg 2020). Although the representations of bidding strength must be distinct from those of dopamine-mediated reinforcement learning credit, they may still involve dopamine, as suggested by the recent finding that in the dorsal tegmentum, dopamine neurons gate associative learning of fear (Groessl *et al.* 2018). The representations of pain in the dopamine system may be acting together with the representations of aversive situations in the hippocampus (Wu *et al.* 2017; Rusu and Pennartz 2020). Furthermore, the confidence differentials among actors may be brought to bear on decision making via a process of drift diffusion (Ratcliff 1981), which has recently been incorporated into a model of reinforcement learning (Pedersen *et al.* 2017).

If, as we hypothesize, bidding costs are indeed represented in the basal ganglia/thalamus/cortex loop, the functional links between this circuit and the many other parts of the brain involved in the pain experience (Chudler and Dong 1995) should be investigated in depth. Of particular interest would be the links to the cingulate cortex, where a lesion can abolish the hurtfulness of pain (Kozuch 2020, p.695), and to the anterior insula, whose normal functioning is critical for the feeling of pain (Gerrans 2020). Interestingly, functional imaging studies in humans indicate that noninvasive therapy methods effectively alter the connectivity between some of the pain-relevant regions; see Lazaridou *et al.* (2017) and Nascimento *et al.* (2018), where the effects on the basal ganglia functional connectivity are explicitly discussed.

We conjecture that it should be possible to use computational means for reconstructing state-space dynamics from EEG or other measurements of brain activity (Moyal *et al.* 2020, 3.1.1) to try and identify the dynamic signature of phenomenal pain.

Further, one may try to modify the relevant dynamics, either by pharmacological means, including common painkillers such as acetaminophen (Durso *et al.* 2015), or by means of therapy (Nascimento *et al.* 2018).

Raising our sights from the implementation- to computational-level connections to other work, we single out for follow-up the predictive processing accounts of pain, which we mentioned in Levels of understanding section. Because the accounts of pain in terms of prediction error dynamics (Fernandez Velasco and Loev 2020) address different levels of explanation compared both to the implementation-level dynamics and to the present paper's functional-evolutionary hypothesis, they may all complement each other. Figuring out whether they do so is a promising direction for future research.

Another computational-level follow-up study could explore a central prediction of our hypothesis, namely, that implementing pain-like costs in an artificial actor-critic system should improve performance, compared to standard reinforcement learning. Here, however, care must be exercised to ensure that this does not result in anything like artificial pain or suffering; Agarwal and Edelman (2020) discuss this issue at some length. The possibility of doing so may depend on whether or not the cost can be put in place without necessarily giving rise to the dynamical signature of pain (see above)—very much an open question.

Finally, we remark that the conceptual roots of the computational mechanism postulated here can be traced back to the early models of parallel distributed processing, implementing functions such as decision-making (Ratcliff 1981), sorting (Dewdney 1984), and salience-driven attention (Koch and Ullman 1985). It would be interesting to bring those ideas up to date, perhaps integrating them with recent work in genetic algorithms (Faliszewski *et al.* 2017) and, of course, reinforcement learning, in particular in single-critic multiple-agent architectures (Zhou *et al.* 2020).

Summary

To understand why pain hurts, one must consider carefully the relationship between its past and future aspects, as well as the relationship between pain and reward. By analogy to the distinction between retributive and consequentialist approaches to justice (e.g., Greene and Cohen 2004; Parfit 2011), it would be useless for pain to be merely a punishment for a poor choice made in the past: it must somehow affect future behavior. Moreover, pain cannot be merely another informational aspect of the process of credit assignment for the purposes of learning. On the account proposed here, pain is functionally orthogonal to such outcome-dependent rewards: it is the cost of bidding on participation in action, whose evolutionary role stems from the need to ensure the reliability of the bids.

Specifically, in vertebrates this functionality is implemented by a distributed network of "actors" that receive performance feedback from a "critic" mechanism. To the extent that actors compete against each other for access to behavioral control, there is a need for a mechanism that would promote honest interactions within the action selection and learning system. Affective experiences reflect the transactions that promote honesty—and thereby make learning more effective—by imposing up-front cost on the actors. In other words, pain hurts because hurting helps keep the actors honest. The price paid in the currency of pain serves to elevate those actors that can afford it and that are therefore more likely to propel the system that embodies them "through the hardships to the stars"—*per aspera ad astra*.

References

Agarwal A, Edelman S. Functionally effective conscious AI without suffering. *J Artif Intell Conscious* 2020;**7**:39–50.

Alexander GE, DeLong MR, Strick PL. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu Rev Neurosci* 1986;**9**:357–81.

Anderson J. Damasio's body-map-based view, Panksepp's affect-centric view, and the evolutionary advantages of consciousness. *South Afr J Philos* 2019;**38**:419–32.

Arbilly M, Lotem A. Constructive anthropomorphism: a functional evolutionary approach to the study of human-like cognitive mechanisms in animals. *Proc R Soc Lond B* 2017;**284**: 20171616.

Atmanspacher H. Dual-aspect monism à la Pauli and Jung. *J Conscious Stud* 2012;**19**:96–120.

Atmanspacher H. Contextual emergence in decompositional dual-aspect monism. *Mind Matter* 2017;**15**:111–29.

Auvray M, Myin E, Spence C. The sensory-discriminative and affective-motivational aspects of pain. *Neurosci Biobehav Rev* 2010;**34**:214–23.

Azqueta-Gavaldon M, Schulte-Göcking H, Storz C, *et al*. Basal ganglia dysfunction in complex regional pain syndrome — a valid hypothesis? *Eur J Pain* 2017;**21**:415–24.

Badre D, Nee DE. Frontal cortex and the hierarchical control of behavior. *Trends Cogn Sci* 2018;**22**:170–88.

Baldassarre, G. and Mirolli, M., editors. *Intrinsically Motivated Learning in Natural and Artificial Systems*. Berlin: Springer, 2013.

Barto AG. Intrinsic motivation and reinforcement learning. In Baldassarre G and Mirolli M (eds), *Intrinsically Motivated Learning in Natural and Artificial Systems*. Berlin: Springer, 2013, 16–47.

Bateson P, Laland KN. Tinbergen's four questions: an appreciation and an update. *Trends Ecol Evol* 2013;**28**:712–8.

Berthier M, Starkstein S, Leiguarda R. Asymbolia for pain: a sensory-lmbic disconnection syndrome. *Ann Neurol* 1988;**24**:41–9.

Blanke O, Metzinger T. Full-body illusions and minimal phenomenal selfhood. *Trends Cogn Sci* 2009;**13**:7–13.

Borsook D, Upadhyay J, Chudler EH, *et al*. A key role of the basal ganglia in pain and analgesia – insights gained through human functional imaging. *Mol Pain* 2010;**6**:27.

Carruthers P. Valence and value. *Philos Phenomenol Res, XCVII* 2018;658–80..

Chen R, Goldberg JH. Actor-critic reinforcement learning in the songbird. *Curr Opin Neurobiol* 2020;**65**:1–9.

Chudler EH, Dong WK. The role of the basal ganglia in nociception and pain. *Pain* 1995;**60**:3–38.

Cleeremans A, Achoui D, Beauny A, *et al*. Learning to be conscious. *Trends Cogn Sci* 2020;**24**:112–23.

Cotterill RMJ. Cooperation of the basal ganglia, cerebellum, sensory cerebrum and hippocampus: possible implications for

cognition, consciousness, intelligence and creativity. *Progr Neurobiol* 2001;**64**:1–33.

Craig AD. How do you feel? Interoception: the sense of the physiological condition of the body. *Nat Rev Neurosci* 2002;**3**:655–66.

Craig AD. A new view of pain as a homeostatic emotion. *Trends Neurosci* 2005;**26**:303–7.

Damasio A, Carvalho GB. The nature of feelings: evolutionary and neurobiological origins. *Nat Rev Neurosci* 2013;**14**:143–52.

Dewdney AK. Computer recreations: on the spaghetti computer and other analog gadgets for problem solving. *Sci Am* 1984;**250**:19–26.

Durso GRO, Luttrell A, Way BM. Over-the-counter relief from pains and pleasures alike: acetaminophen blunts evaluation sensitivity to both negative and positive stimuli. *Psychol Sci* 2015;**26**:750–8.

Edelman GM, Finkel L. Neuronal group selection in the cerebral cortex. In Edelman GM, Gall WE, and Cowan WM(eds), *Dynamical Aspects of Neocortical Function*. New York: Wiley, 1984 653–95.

Edelman S. *Computing the Mind: How the Mind Really Works*. New York, NY: Oxford University Press, 2008.

Edelman S. Vision, reanimated and reimagined. *Perception* 2012;**41**:1116–27.

Edelman S, Moyal R, Fekete T. To bee or not to bee? A commentary on insects have the capacity for subjective experience, C. Klein & A. B. Barron.*Animal Sentience* 2016.

Eisenberger NI. The pain of social disconnection: examining the shared neural underpinnings of physical and social pain. *Nat Rev Neurosci* 2012;**13**:421–34.

Faliszewski P, Sawicki J, Schaefer R, *et al*. Multiwinner voting in genetic algorithms. *IEEE Intell Syst* 2017;**32**:40–8.

Fazeli S, Büchel C. Pain-related expectation and prediction error signals in the anterior insula are not related to aversiveness. *J Neurosci* 2018;**38**:6461–74.

Feinberg TE, Mallatt J. The evolutionary and genetic origins of consciousness in the Cambrian period over 500 million years ago. *Front Psychol* 2013;**4**:667.

Fekete T, and, Edelman S. Towards a computational theory of experience. *Conscious Cogn* 2011;**20**:807–27.

Feldman Barrett, L. Psychological construction: the Darwinian approach to the science of emotion. *Emotion Rev* 2013;**5**:379–89.

Fernandez Velasco, P. and Loev, S. Affective experience in the predictive mind: a review and new integrative account. *Synthese* 2020; doi: 10.1007/s11229-020-02755-4.

Fink SB. Independence and connections of pain and suffering. *J Conscious Stud* 2011;**18**:46–66.

Foster NN, Korobkova L, Garcia L, *et al*. The Mouse Cortico-Basal Ganglia-Thalamic Network. bioRxiv preprint, 2020; doi: 10.1101/2020.10.06.326876.

Friend DM, Kravitz AV. Working together: basal ganglia pathways in action selection. *Trends Neurosci* 2014;**37**:301–3.

Friston KJ. The free-energy principle: a unified brain theory? *Nat Neurosci* 2010;**11**:127–38.

Friston KJ, Wiese W, Hobson JA. Sentience and the origins of consciousness: from Cartesian duality to Markovian monism. *Entropy* 2020;**22**:516.

Frith CD, Metzinger T, What's the use of consciousness? How the stab of conscience made us really conscious. In Engel AK, Friston KJ, and Kragic D (eds), *The Pragmatic Turn: Toward Action-Oriented Views in Cognitive Science*, volume **18** of Strüngmann Forum Reports. Cambridge, MA: MIT Press, 2016, 197–224.

Gerrans P. Pain asymbolia as depersonalization for pain experience. An interoceptive active inference account. *Front Psychol* 2020;**11**:523710.

Gershman SJ. Context-dependent learning and causal structure. *Psychon Bull Rev* 2017;**24**:557–65.

Gershman SJ, Uchida N. Believing in dopamine. *Nat Rev Neurosci* 2019;**20**:703–14.

Godfray HCJ. Signalling of need by offspring to their parents. *Nature* 1991;**352**:328–30.

Godfray HCJ, Johnstone RA. Begging and bleating: the evolution of parent–offspring signalling. *Philos Trans R Soc B* 2000;**355**:1581–91.

Grafen A. Biological signals as handicaps. *J Theoret Biol* 1990;**144**:517–46.

Graybiel AM. Habits, rituals, and the evaluative brain. *Annu Rev Neurosci* 2008;**31**:359–87.

Greene JD, Cohen JD. For the law, neuroscience changes nothing and everything. *Philos Trans R Soc Lond B* 2004;**359**:1775–85.

Grillner S, Robertson B. The basal ganglia over 500 million years. *Curr Biol* 2016;**26**:R1088–R1100.

Groessl F, Munsch T, Meis S, *et al*. Dorsal tegmental dopamine neurons gate associative learning of fear. *Nat Neurosci* 2018;**21**:952–62.

Hagiwara K, Garcia-Larrea L, Tremblay L, *et al*. Pain behavior without pain sensation: an epileptic syndrome of "symbolism for pain"? *Pain* 2020;**161**:502–8.

Harris KD, Weiss M, Zahavi A. Why are neurotransmitters neurotoxic? An evolutionary perspective. *F1000Research* 2014;**3**:179.

Higham JP. How does honest costly signaling work? *Behav Ecol* 2014;**25**:8–11.

Hohwy J. New directions in predictive processing. *Mind Lang* 2020;**35**:209–23.

Humphries MD, Prescott TJ. The ventral basal ganglia, a selection mechanism at the crossroads of space, strategy, and reward. *Progr Neurobiol* 2010;**90**:385–417.

Joel D, Weiner I. The organization of the basal ganglia-thalamo-cortical circuits: open interconnected rather than closed segregated. *Neuroscience* 1994;**63**:363–79.

Joffily M, Coricelli G. Emotional valence and the free-energy principle. *PLoS Comput Biol* 2013;**9**:e1003094.

Kanai R, Komura Y, Shipp S, *et al*. Cerebral hierarchies: predictive processing, precision and the pulvinar. *Philos Trans R Soc B* 2015;**370**:20140169.

Kihlstrom JF, The rediscovery of the unconscious. In Morowitz H and Singer J (eds), *The Mind, the Brain, and Complex Adaptive Systems*, volume **22** of Santa Fe Institute Studies in the Sciences of Complexity. Reading, MA: Addison-Wesley, 1995, 123–43.

Kirk, R. Zombies. In Zalta, EN (ed.), *The Stanford Encyclopedia of Philosophy*, Spring 2019 edition. 2019.

Klein C. What pain asymbolia really shows. *Mind* 2015;**124**:493–516.

Koch C, Ullman S. Selecting one among the many: a simple network implementing shifts in selective visual attention. *Hum Neurobiol* 1985;**4**:219–27.

Koechlin E. Prefrontal executive function and adaptive behavior in complex environments. *Curr Opin Neurobiol* 2016;**37**:1–6.

Kozuch B. No pain, no gain (in Darwinian fitness): a representational account of affective experience. *Erkenntnis* 2020;**85**:693–714.

Kuner R, Flor H. Structural plasticity and reorganisation in chronic pain. *Nat Rev Neurosci* 2017;**18**:20–30.

Laidre ME, and, Johnstone RA. Animal signals. *Curr Biol* 2013;**23**:R829–R833.

Laland KN, Sterelny K, Odling-Smee J, *et al*. Cause and effect in biology revisited: is Mayr's proximate-ultimate dichotomy still useful? *Science* 2011;**334**:1512–6.

Lazaridou A, Kim J, Cahalan CM, *et al*. Effects of cognitive-behavioral therapy (CBT) on brain connectivity supporting catastrophizing in fibromyalgia. *Clin J Pain* 2017;215–21.

Levy D, Glimcher P. The root of all value: a neural common currency for choice. *Curr Opin Neurobiol* 2012;**22**:1027–38.

Lindquist KA, Wager TD, Kober H, *et al*. The brain basis of emotion: a meta-analytic review. *Behavioral and Brain Sciences* 2012;**35**:121–202.

Lloyd K, Dayan P. Safety out of control: dopamine and defence. *Behav Brain Funct* 2016;**12**:15.

Marr D, Poggio T. From understanding computation to understanding neural circuitry. *Neurosci Res Prog Bull* 1977;**15**:470–88.

Márton M. What does the zombie argument prove? *Acta Anal* 2019;**34**:271–80.

Mayr E. Cause and effect in biology. *Science* 1961;**134**:1501–6.

Merker B. Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. *Behav Brain Sci* 2007;**30**:63–81.

Metzinger T. *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, MA: MIT Press, 2003.

Metzinger T. Suffering, the cognitive scotoma. In Almqvist K. and Haag A (eds) *The Return of Consciousness*. Stockholm: Axel and Margaret Ax:son Johnson Foundation, 2017, 237–62.

Metzinger T. Minimal phenomenal experience, 2018. Third Carnap Lecture, given on 9 March,2018, at the Ruhr-Universität Bochum.

Metzinger T. Minimal phenomenal experience: meditation, tonic alertness, and the phenomenology of "pure" consciousness. *Philos Mind Sci* 2020;**1**:7–44.

Minsky M. Steps toward artificial intelligence. *Proc Inst Radio Eng* 1961;**49**:8–30.

Moyal R, Fekete T, and Edelman S. Dynamical Emergence Theory (DET): a computational account of phenomenal consciousness. *Minds Mach* 2020;**30**:1–21.

Nascimento SS, Oliveira LR, DeSantana JM. Correlations between brain changes and pain management after cognitive and meditative therapies: a systematic review of neuroimaging studies. *Complement Ther Med* 2018;**39**:137–45.

Nesse RM, Schulkin J. An evolutionary medicine perspective on pain and its disorders. *Philos Trans R Soc B* 2019;**374**:20190288.

Oizumi M, Albantakis L, Tononi G. From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Comput Biol* 2014;**10**:e1003588.

O'Reilly RC. Biologically based computational models of high-level cognition. *Science* 2006;**314**:91–4.

Panksepp J. The neuro-evolutionary cusp between emotions and cognitions: implications for understanding consciousness and the emergence of a unified mind science. *Evol Cogn* 2001;**7**:141–63.

Panksepp J. Affective consciousness: core emotional feelings in animals and humans. *Conscious Cogn* 2005;**14**:30–80.

Parfit D. *On What Matters*. Oxford, UK: Oxford University Press, 2011.

Pedersen ML, Frank MJ, Biele G. The drift diffusion model as the choice rule in reinforcement learning. *Psychon Bull Rev* 2017;**24**:1234–51.

Pierce JE, Péron J. The basal ganglia and the cerebellum in human emotion. *Soc Cogn Affect Neurosci* 2020;599–613.

Ploner M, Freund HJ, Schnitzler A. Pain affect without pain sensation in a patient with a postcentral lesion. *Pain* 1999;**81**:211–4.

Polcyn K. Does the conceivability of zombies entail their possibility? *Organon F* 2020;**3**:395–410.

Ratcliff R, Parallel processing mechanisms and processing of organized information in human memory. In Anderson JA and Hinton GE (eds), *Parallel Models of Associative Memory*. Hillsdale, NJ: Erlbaum, 1981.

Redgrave P, Rodriguez M, Smith Y, *et al*. Goal-directed and habitual control in the basal ganglia: implications for Parkinson's disease. *Nat Rev Neurosci* 2010;**11**:760–72.

Redgrave P, Vautrelle N, Reynolds JNJ. Functional properties of the basal ganglia's re-entrant loop architecture: selection and reinforcement. *Neuroscience* 2011;**198**:138–51.

Redgrave P, Gurney K, Stafford T, *et al*. The role of the basal ganglia in discovering novel actions. In Baldassarre G and Mirolli M (eds), *Intrinsically Motivated Learning in Natural and Artificial Systems*. Berlin: Springer, 2013, 129–50.

Rodríguez-Gironés MA, Cotton PA, Kacelnik A. The evolution of begging: signaling and sibling competition. *Proc Natl Acad Sci USA* 1996;**93**:14637–41.

Rusu SI, Pennartz CMA. Learning, memory and consolidation mechanisms for behavioral control in hierarchically organized cortico-basal ganglia systems. *Hippocampus* 2020;**30**:73–98.

Saga Y, Ruff CC, Tremblay L. Disturbance of approach-avoidance behaviors in non-human primates by stimulation of the limbic territories of basal ganglia and anterior insula. *Eur J Neurosci* 2019;**49**:687–700.

Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw* 2015;**61**:85–117.

Tinbergen N. On aims and methods in ethology. *Zeitschrift Für Tierpsychologie* 1963;**20**:410–33.

Tononi G. Consciousness as integrated information: a provisional manifesto. *Biol Bull* 2008;**215**:216–42.

Van de Cruys S, Affective value in the predictive mind. In Metzinger T and Wiese W (eds), *Philosophy and Predictive Processing*, Chapter 24. Frankfurt am Main: MIND Group, 2017. doi: 10.15502/9783958573253.

Walters ET, Williams ACC. Evolution of mechanisms and behaviour important for pain. *Philos Trans R Soc B* 2019;**374**:20190275.

Wickens JR. Synaptic plasticity in the basal ganglia. *Behav Brain Res* 2009;**199**:119–28.

Windt JM. Just in time — dreamless sleep experience as pure subjective temporality. In Metzinger TK and Windt JM (eds), *Open MIND*. Frankfurt am Main: MIND Group, 2015.

Worley S. Conceivability, possibility and physicalism. *Analysis* 2003;**63**:15–23.

Wu C-T, Haggerty D, Kemere C, *et al*. Hippocampal awake replay in fear memory retrieval. *Nat Neurosci* 2017;**20**:571–83.

Zahavi A. Mate selection — a selection for a handicap. *J Theoret Biol* 1975;**53**:205–14.

Zahavi A. Animal communication: the handicap principle. In Callan H (ed.), *The International Encyclopedia of Anthropology*. New York, NY: John Wiley & Sons, 2018. doi: 10.1002/9781118924396.wbiea1804.

Zahavi A and Zahavi A. *The Handicap Principle: A Missing Piece of Darwin's Puzzle*. New York, NY: Oxford University Press, 1997.

Zahavi A, and, Zahavi A. The logic of analog signaling and the theory of signal selection. *Israel J Ecol Evol* 2012;**58**:269–78.

Zhou M, Liu Z, Sui P, *et al*. In Balcan M and Lin H (eds), *Advances in Neural Information Processing Systems*, vol. 33 (NeurIPS 2020). 2020.