NEUROSCIENCE of Consciousness

Neuroscience of Consciousness, 2024, **2024(1)**, niae034

DOI: https://doi.org/10.1093/nc/niae034

Rapid Communication

Making sense of feelings

Brian Key[1,*] and Deborah J. Brown[2,*]

[1]School of Biomedical Sciences, University of Queensland, Brisbane, Queensland 4072, Australia
[2]School of Historical and Philosophical Inquiry, University of Queensland, Brisbane, Queensland 4072, Australia

*Corresponding author. School of Biomedical Sciences, University of Queensland, Brisbane 4072, Australia. E-mail: brian.key@uq.edu.au; School of Historical and Philosophical Inquiry, University of Queensland, Brisbane, Queensland 4072, Australia. E-mail: deborah.brown@uq.edu.au

Abstract

Internal feeling states such as pain, hunger, and thirst are widely assumed to be drivers of behaviours essential for homeostasis and animal survival. Call this the 'causal assumption'. It is becoming increasingly apparent that the causal assumption is incompatible with the standard view of motor action in neuroscience. While there is a well-known explanatory gap between neural activity and feelings, there is also a disjuncture in the reverse direction—what role, if any, do feelings play in animals if not to cause behaviour? To deny that feelings cause behaviours might thus seem to presage epiphenomenalism—the idea that subjective experiences, including feelings, are inert, emergent and, on some views, non-physical properties of brain processes. Since epiphenomenalism is antagonistic to fundamental commitments of evolutionary biology, the view developed here challenges the standard view about the function of feelings without denying that feelings have a function. Instead, we introduce the 'sense making sense' hypothesis—the idea that the function of subjective experience is not to cause behaviour, but to explain, in a restricted but still useful sense of 'explanation'. A plausible framework is derived that integrates commonly accepted neural computations to blend motor control, feelings, and explanatory processes to make sense of the way feelings are integrated into our sense of how and why we do and what we do.

Keywords: subjective experience; phenomenal consciousness; qualia; awareness; neural circuitry

Introducing the sense making sense hypothesis

I remove my hand from the hotplate because it hurts. The mouse eats the cheese because she is hungry. What could be more obvious than that it is the feeling—pain and hunger, respectively—that is the cause of each subsequent behaviour? The negative or positive valences of feelings seem intuitively to act as primary drives for survival. We call this commonsense idea that feelings cause behaviour the 'causal assumption'. The academic literature is replete with the standard view that feelings seem to inform our decision-making and acting because of this assumption. For example, Wiech and Tracey (2013) claim that 'pain signals potential harm to the organism, it immediately attracts attention and motivates decisions and action'. In discussing neural circuits for motivational states, Lee and Wu (2020) stress that 'the need for nutrients generates hunger, which serves as the motivational drive for eating'. Similarly, 'a pain in your hand seems to motivate actions such as withdrawing from a painful stimulus'; it is a 'direct motivational force' that 'tell(s) us how to act' (McClelland and Jorba 2023).

From a neuroscientific perspective, it can seem completely mysterious how something as ineffable as a feeling could be causally efficacious over physical events, such as action potentials in neurons, leading some to conclude that feelings should instead be considered epiphenomena without any function (Robinson 2010). But such reactions are difficult to reconcile with the widely accepted idea that 'nothing makes sense except in the light of evolution' (Dobzhansky 1973). Why should the brain evolve mechanisms to produce conscious feelings that demand considerable energy expenditure (Mashour et al. 2020), if feelings were causally redundant? If the causal assumption is false, an alternative and plausible explanation of the function of feelings is required, and one that preferably explains why the causal assumption is so intuitively appealing.

Here, we propose a different account of the function of feelings, which we call the 'sense making sense hypothesis' (SMS), where 'sense' refers broadly to feeling states and 'making sense' to an explanatory function. According to SMS, the primary function of feelings is not to cause behaviour but to 'explain' it. This view draws inspiration from Huxley who long ago argued that 'the feeling we call volition is not the cause of a voluntary act, but the symbol of that state of the brain which is the immediate cause of that act' (Huxley 1874). We extend Huxley's view by suggesting that valenced feelings are 'symbols' that refer to the neural

Received 10 March 2024; revised 12 August 2024; accepted 27 August 2024

causes of behaviour. The fact that the real cause of behaviour—the nonconscious neural processing—is a common cause of both the behaviour and the feeling is what explains why feelings seem not only to control behaviour but also lends insight into the proper function of feelings. Viewed in this way, feelings provide subjective awareness of the cause of behaviour rather than being causes themselves. The SMS hypothesis is grounded in the premise that only complex brains (*i.e.* brains that can re-represent neural information at different levels of abstraction; Key et al. 2022) that have evolved information-seeking neural circuits capable of inferring (*i.e.* predicting) their own operations have the capacity for feeling.

The SMS hypothesis proposes that the feeling of hunger is the mouse's conscious awareness or prediction of what is causing it to search for food—*i.e.* the feeling explains its behaviour by referring to its predicted neural cause. It is generally accepted that the brain has no direct access to knowledge of the world and uses inferences by internal neural models to explain its external interactions (Helmholtz 1925; see discussion in the following section). Given that subjective experience is an awareness of those predictive neural processes, we contend that feelings such as hunger are the conscious brain's way of explaining to itself why it is foraging. The SMS hypothesis proposes that the pain humans experience after touching a hotplate automatically or implicitly explains hand withdrawal—it clearly not causally necessary for the behaviour because it is reflexive and can be executed without conscious awareness (Riddoch 1917) and there is good evidence that pain proceeds rather than precedes behaviour (Campbell and LaMotte 1983, Thorell et al. 2023). Moreover, the brain can learn to inhibit the neural processing causing such behaviour and, hence, pain is not experienced, when, for example, we nurse a fractured arm to stop movement. In this case, pain is not causing the guarding behaviour but rather it is the anticipation or prediction that inhibiting the neural cause will prevent the injury-inducing movements that is the driving cause. The guarding behaviour is ultimately the result of hierarchically controlled neural circuits executing commands driven by goal-directed brain processes (Ashe and Georgopoulos 1994, Kalaska et al. 1997, Buneo and Andersen 2006). This dissociation between feeling state and action is like that proposed between the sense of voluntary control of movement and the neural circuits controlling motor actions (Haggard 2005, 2008, Soon et al. 2008).

Our task here in the following sections is 3-fold: first, to clarify the nature of the explanatory function of feelings; second, to outline a biologically plausible neural framework for SMS; and third, to highlight some of the benefits of the SMS view. It is not our intention to address how brain circuits generate feelings. For our purposes, it is enough to accept that feelings are conscious and whether that is because they are active in some sort of conscious space, a 'global workspace' (Dehaene and Naccache 2001), or a higher-order centre (Brown et al. 2019, Lau et al. 2022) remains to be empirically determined.

What kind of explanation could a feeling be?

The SMS hypothesis advances the idea that valenced feelings are not causative of action and are instead informative by explaining or making sense of one's behaviour. What then do we mean when we say that feelings 'explain' or to 'make sense' of behaviour?

The notion of 'explanation' we use is quite specific. What makes how a feeling 'makes sense' an explanation is its inferential aspect. In 1910, Helmholtz proposed that the brain is primarily engaged in inference since its only interaction with the external world is via internal neural activity (Helmholtz 1925). This idea, still widely accepted in neuroscience, has progressed into predictive coding models of the brain (Friston et al. 2006, Hohwy 2013), according to which the brain infers the nature of sensory stimuli using top-down, nonconscious predictions produced by internal 'generative' models of the environment—generative in the sense of predicting the cause (or input) of an output. These predictions are inferences to the 'best explanation' of a cause, which supports the claim that a fundamental function of the brain is to 'explain' its sensory and motor interactions with the world. The SMS hypothesis proposes that feelings embody predictions of internal models that infer the cause of behaviour. While feelings such as pain may represent (or relate to) the nature of the sensory stimulus (*e.g.* sharp, dull, cold, hot), it is the unpleasant quality of the experience that explains (refers to or accounts for) the cause of the behaviour.

For those who conflate 'cause' and 'explanation' and think of causal relations as necessary to ground explanations and predictions in science, SMS is bound to disappoint. But there is no necessity here. While explanations may cite causes, they need not themselves be causes. It is not by anyone's explaining the volcano's eruption that the eruption is caused. And some explanations, *i.e.* false ones, can have effects (*e.g.* on what people believe) without citing any causes at all. We thus resist the conflation of 'cause' and 'explanation'. While explanations may cite causes, they do not in virtue of that alone become causes themselves.

One might object that if feelings give rise to explanations, then they are causes of mental behaviours rather than motor acts. Viewed in this way, pain generates a propositional explanation of behaviour. The SMS hypothesis instead considers that a feeling such as pain does not lead to an explanation but 'constitutes' the explanation. A precursor to this idea can be found in Russell's (1910) referring to subjective experience (*i.e.* feelings) as a kind of 'knowledge by acquaintance' (although we distance ourselves from the idea that feelings acquaint us with Platonic entities like sense data). The main point is to recognize that not all explanations need to consist of propositional knowledge, which would depend on more complex cognitive processing. Knowledge by acquaintance is a form of direct awareness (see also Wegner's idea of direct knowledge discussed further). More recently, Horgan and Kriegel (2007) have claimed that 'when a phenomenally conscious state represents something, it makes the subject aware of what it represents'. According to SMS, feelings refer to or 'stand in for' the causes of behaviour in subsequent neural processing. We further suggest that the conscious awareness is implicit, following Horgan and Kriegel's (2007) contention that awareness is 'inbuilt' or inherent in the experience. There is no need for an extra mental step to awareness as what is represented as awareness is 'a component of the experience itself'. Interestingly, Levine (2019) also proposes that knowledge by acquaintance has no existence outside of what he calls the 'virtual world' of conscious experience, which, on our view, confirms their lack of causal efficacy with respect to behaviour.

Giustina (2022) points out a common confusion in the literature about knowledge by acquaintance, which is relevant for our purposes. Too often it is assumed incorrectly that knowledge of perceptual states is caused by the states themselves. Giustina indicates that Russell's original intention was that knowledge by acquaintance be constituted by acquaintance and not caused by it. Using pain as an example of knowledge by acquaintance, Guistina highlights how the experience of pain is only known

by experience—*i.e.* by direct acquaintance. In line with this suggestion, SMS proposes that feelings explain behaviour because they constitute direct, non-propositional, experiential knowledge of internal brain states (*i.e.* predictions) that stand for the cause of behaviour.

The idea that the human brain is inclined to generate implicit explanations of behaviour is supported empirically. Feelings have an automatic explanatory function that can be revealed in some experimentally produced sensory illusions. Take as an example, the cutaneous rabbit effect (Geldard and Sherrick 1972). This illusion involves the delivery of five brief pressure pulses at three locations on the forearm: first at the wrist, next at the middle of the forearm, and the last near the elbow. Surprisingly, the subject experiences these pulses not as occurring at three separate sites but instead as occurring at 15 evenly spaced taps sequentially moving along the length of forearm—as if a rabbit were hopping from the wrist up to the elbow. The remarkable feature of this illusion is that the brain implicitly infers that discrete and spatially segregated taps on the forearm are not a typical natural occurrence and that a better explanation (*i.e.* representation) would be to experience them as if something were moving up the forearm. These inferences can be replicated in computational models based on generating predictions (Goldreich and Tong 2013). Two take-home messages are worth noting here. First, it is the predictions of brain models and not the direct sensory inputs that are perceived as feelings—*i.e.* feelings are an automatic reference to predicted causes of behaviour not the actual causes of behaviour. Second, feelings are implicitly used to explain how stimuli should be experienced, which is most likely to be highly advantageous when sensory information is so often both inherently noisy and novel (Faisal et al. 2008).

Pioneering investigations by Gazzaniga and LeDoux with split brain patients demonstrate how subjects are also highly motivated to offer cognitively based explanations of their behaviour that are, nonetheless, false causal narratives. Among a series of elegant psychobiology experiments, Gazzaniga and LeDoux tested a patient (called 'P.S.') who could read from both sides of the brain (Gazzaniga and LeDoux 1978, p. 86). When instructions to perform actions were projected only to the right brain, P.S. acted accordingly. When the speaking and unaware left brain was then asked why they performed these behaviours, P.S. replied with fictitious explanations. For example, P.S. stood up when the command to stand was shown to the right brain. The left brain then explained their behaviour by saying that they needed to stretch. Gazzaniga (1995) argued that the conscious brain was driven to provide explanations of behaviour, coining the term 'left hemisphere interpreter'. Remarkably, these cognitive-based explanations are also employed to account for feelings which are artificially generated by direct electrical stimulation of the cortex. Stimulation of the anterior supplementary motor area induces laughter and a sensation of mirth in an awake patient undergoing brain surgery (Fried et al. 1998). When queried as to why they laughed, the patient retorted with a fictious story about something funny in the operating theatre. These case studies reveal how subjects willingly fabricate events to satisfy a desire to explain their behaviour. They suggest that the need to explain ourselves to ourselves and others runs deep and that we will, if we must, fabricate events just to supply an explanation. The explanatory function of feelings appears to have both an essential cognitive function and play an important role in our social interactions (Malle 2006).

A conceptual framework for the SMS hypothesis

It is our contention that where the informational-seeking algorithms executed by brain circuits reach a point of computational complexity that enable them to explain to themselves their own operations, that explanation is consciously presented in its most basic form as a feeling state. This basic premise has similarities to another recent described framework of consciousness referred to as the self-organizing meta-representational account (SOMA) (Cleeremans et al. 2020). SOMA suggests that feelings are not intrinsic properties of neural activity but rather the outcome of nonconscious plasticity mechanisms that enables the brain to learn to re-represent its own internal activity as meta-representations. In this section, we map out our conceptual framework for the SMS hypothesis, which builds on this important idea that learning is foundational for building models, the predictive outputs of which represent the causes of behaviour, and which form the basis of feeling states such as pain.

Our conceptual framework for the explanatory nature of feelings is based on two common predictive models: forward models and inverse models (Fig. 1). For simplicity, the operations of a functional system in the human brain can be reduced to a linear series of three modules consisting of 'input', 'processing', and 'output' modules. In this context, a forward model is a neural network that learns to predict the output of the processing stream given some input. The model is trained using feedback (called the 'error signal') obtained by comparing the real output of the processing stream with the predicted output. In contrast, an inverse model takes the current output of the system and predicts what input most likely generated it. A typical example of the combined use of forward and inverse models in physical systems is weather forecasting. Given current climatic conditions, a forward model predicts the future state of the weather. The accuracy of the forward prediction depends on knowing the likely causes of the current conditions as determined by an inverse model. In the mammalian nervous system, combinations of forward and inverse models have been suggested to underpin processes such as motor control, visual perception, action feeling states, and metacognition (Kawato et al. 1993, Pacherie 2008, Kawato and Cortese 2021, Cortese and Kawato 2024).

We propose that for feelings to have an explanatory role in behaviour the brain needs to have information about the internal cause of the behaviour because this cause constitutes the explanation of its behaviour. Using the example of pain, we contend that the internal cause of behaviour is the driving input that creates the motor command and generates the behaviour. This input comprises integrated information from sensory systems and internal brain states (*e.g.* nociception, proprioception, somatosensation, vision, attention, and background activity). The brain lacks direct access to the causal role played by each of these components in the driving input. However, the motor command is available for inferring the nature of the relevant inputs. By inputting a copy of the motor command into an inverse model, it is possible to generate a prediction of the driving input and hence the cause of the behaviour. However, this approach is too slow and would lead to the cause being determined after the motor command and behaviour. An alternative strategy is to use a forward model to first produce a rapid prediction of the future motor command based on a copy of the driving inputs before the real motor command has been generated. This predicted motor command is then

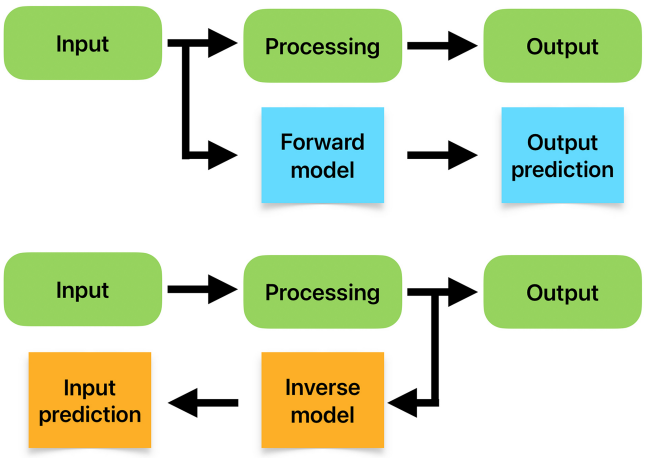

Figure 1 Diagrammatic representation of an input-output brain pathway consisting of an input, a processing module and an output. In the upper pathway, a copy of the input is sent to a forward model that generates a rapid prediction of the output. In the lower pathway, a copy of the output is sent to an inverse model that rapidly predicts the input. The forward and inverse models are trained using an error signal generated by comparing the output prediction and input prediction with the real output or input, respectively

inputted into an inverse model to predict the likely driving input—*i.e.* the inferred cause. This inferred cause is further processed and subjectively experienced, *e.g.* as pain. Therefore, a feeling like pain is the explanation to the conscious brain of what caused the 'pain behaviour'. Further, we present the framework in more detail together with some evidence for the plausible cortical location of the neural circuits executing these computations.

Our framework for the SMS hypothesis (Fig. 2) is hierarchical with a base layer generating motor commands leading to behaviour. This layer involves nonconscious processing, which allows many behaviours to be executed without conscious control (McBride et al. 2012). The premotor cortex is likely an important component in this layer since direct electrical stimulation here in patients undergoing brain surgery elicits motor actions in the absence of either a subjective desire to act or of the movement having been performed (Desmurget et al. 2009). A copy of the driving input is relayed to the next layer network, which contains a forward model. This model outputs a prediction of the base layer's motor command. This predicted motor command allows for the generation of upstream feelings that are separate and hence dissociable from the nonconsciously controlled behaviour. In the next layer, the predicted motor command is then fed into separate inverse and forward models. The inverse model generates a prediction of the cause of the predicted motor command. This predicted cause is the predicted driving input which is then fed back into the base layer to drive the behaviour. The predicted cause is also simultaneously processed by neural circuits to generate the conscious desire to act.

The predicted motor command also enters a second forward model, which generates the predicted sensory output as if a movement were to occur. This predicted sensory feedback is then fed back into the driving input to provide rapid feedback to the generation of the motor command. This predicted sensory output is consciously experienced as a feeling of awareness of having seemingly performed a movement. The second layer containing the inverse and forward model processing the predicted motor command is likely located in the posterior parietal cortex. There is evidence in monkeys that the lateral intraparietal area of the posterior parietal contains neurons that respond to future sensory stimuli even before movements are executed (Duhamel et al. 1992). This suggests that this region generates sensory predictions from forward models (Mulliken et al. 2008, Medendorp and Heed 2019).

Many different states of awareness have been proposed under the broad banner of the phenomenology of action (Pacherie 2008). We are interested here in those states that have an associated feeling, albeit of neutral valence, because they can provide insights into the brain regions capable of generating feelings. Of relevance is the feeling of the desire to act (also called the intention to act, the thought to act, or volition), the feeling of initiating an action, the feeling of agency (sense of being an author of one's action), and, finally, the awareness of movement (Haggard 2005, Pacherie 2008, Amanzio et al. 2010, Darby et al. 2018). What is important here is not so much the nomenclature but the fact that for each of these states there is something that it feels like to be in them. The ability of subjects to report such states has enabled the localization of these feelings in the brain in both experimental and clinical studies.

Weak direct electrical stimulation of the human posterior parietal cortex elicits a felt desire to act, while strong direct electrical stimulation creates both this desire as well as an awareness of movement, despite the absence of movement (Desmurget et al. 2009). This awareness of movement without any accompanying movement is consistent with the increased speed of processing in predictive models in comparison to the base layer and it allows adjustment to be made to movements before their completion. Desmurget et al. (2018) and Fornia et al. (2022) have more recently shown that direct electrical stimulation of the superior parietal lobule and intraparietal cortex (positioned between the superior and inferior parietal lobules) in the posterior parietal cortex neither generates movements, nor creates any desire to move, nor leads to any sensation of having performed a movement. These results suggest that the inverse and forward models of layer 2 are both present in the inferior parietal lobule of the posterior parietal cortex. Interestingly, stimulation of the superior parietal lobule and intraparietal cortex prevents patients from initiating voluntary hand movements and stops any ongoing hand movements. This inhibitory action occurs independently of any voluntary intention to stop movements, which demonstrates the separation of neural circuitry mediating feelings to act from the execution of the behaviour. This dissociation of functions is consistent with the earlier report by Desmurget et al. (2009), which suggests that the desire to act or volition is not necessarily causative, as our intuitions would lead us to believe (Haggard 2008, Fried et al. 2017, Wegner 2017).

We are not wedded to the precise location of the forward and inverse models given that other studies have proposed alternative regions (Garbarini et al. 2019, Bruno et al. 2023). Our framework is focused on how subjective experience lacks casual power over behaviour which is consistent with Desmurget et al. (2009) and leaves the location of the cortical regions executing specific functions to be determined empirically. Nonetheless, there is converging evidence that activity centred around or near the intraparietal sulcus in humans is responsible for many somatosensory perceptions (Pereira et al. 2021, De Havas et al. 2022). It is important to note that the posterior parietal cortex is well known to be involved in generating internal models of the environment and state prediction errors (Gläscher et al. 2010), but this pertains to conscious rather than nonconscious processing.

The SMS hypothesis may at first glance seem at odds with the generally acknowledged causative role that feelings play

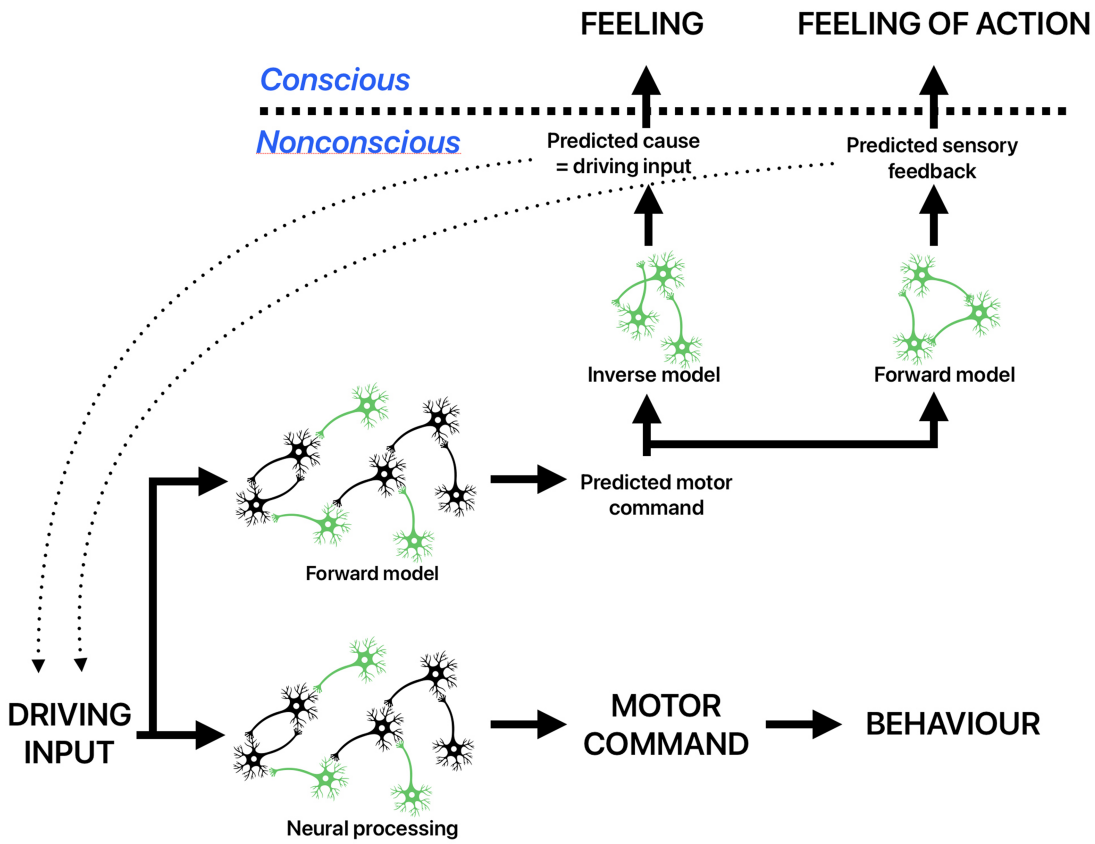

Figure 2 Feelings explain behaviour framework. The underlying premise here is that the conscious brain seeks to explain to itself why it acts in the way it does. A copy of the driving input is sent to a forward model, which predicts the motor command. This predicted motor command is simultaneously entered into both an inverse model and a second forward model. The inverse model outputs the cause of the predicted motor command, which is equivalent to the prediction of the driving input. The second forward model outputs the effect of the motor command which is the predicted sensory feedback arising from the behaviour. Both the predicted driving input and predicted sensory feedback are generated prior to the behaviour and can be fed back into the driving input to rapidly adjust the motor command in real time. The predicted driving input and sensory feedback are what become experienced as conscious feelings. Notably, it is the predicted sensory feedback that contributes specifically to feelings of action. The square dotted line separating conscious and nonconscious processing is illustrative and is not intended to represent the existence of a conscious place in the brain

in reasoning and decision-making (Lombrozo 2006, Malezieux et al. 2023). A large literature supports affective decision-making and the cognitive benefits of feelings-based emotional reasoning (Zajonc 1980, Oatley and Johnson-Laird 1987, Han et al. 2007, Quartz 2009, Todd et al. 2020, Malezieux et al. 2023). Much of this research presupposes the causal assumption that feelings mediate and motivate behaviour and action (Oatley and Johnson-Laird 2011). The SMS hypothesis provides a different way of thinking about these colloquially termed 'gut feelings'. It suggests that feelings are the explanations of the decisions made rather than the cause of actions, the causes more likely being underlying nonconscious neural computations.

Interestingly, Pacherie (2008) has developed a framework using forward and inverse models to explain how feeling states of action could arise during motor control. While this approach demonstrates the power of using pairs of these models in series, it is conceptually quite different to our framework. Pacherie uses these models to describe how feelings are generated, whereas we instead accept the presence of feelings and show how feelings come to represent causes of behaviour. In our case, the forward and inverse models are operational at nonconscious levels and the outputs of an inverse model are what are subsequently processed to generate feelings. In contrast, Pacherie proposes that the direct outputs of forward and inverse models are feeling states associated with action. For example, the feeling of being aware of what we are doing involves the predictions of both an inverse model that is constitutive of our immediate conscious goals as well as a forward model that predicts a future sensory state. While feelings are attributed to model predictions, the explanatory gap between subjective states and neural processing in these models remains.

Our view that feelings do not cause pain behaviour is consistent with the proposal that conscious thoughts do not cause voluntary action as outlined in a simple model of a mental system proposed by Wegner (2017, p. 63). Wegner (2004) attempts to integrate conscious will and valenced feelings in a concept he refers to as "emotion of authorship". In drawing this relationship, Wegner suggested that the conscious will can somehow automatically know what the body is doing; people know "by the sheer quality of the experience just what happened" (Wegner 2004, p. 658). Here, valenced feelings seem to carry some special knowledge of action and hence, are informative (Wegner 2017, p. 309). He considers that emotions are plausible explanations of the likely causes of behaviours (without being the actual causes), such as when a person acts "out of blind rage or profound sadness" (Wegner 2017, pp. 84–86). If conscious willing is like an emotional feeling, then that also should be informative—but informative of what? Wegner's answer is that the feeling of conscious volition functions in "making sense" of the cause of one's behaviour among all possible candidate causes (Wegner 2017, pp. 312–14). While Wegner advocates that a conscious willing is not causative of actions, he

believes that the feeling of it is somehow causative at the level of psychological states. That is, the feeling of a conscious volition provides an individual with a sense of perceived or illusory control, which is ultimately needed for psychological well-being. He is admitting here that while feelings may not cause actions, they do cause other feelings (*i.e.* feelings of well-being), and thus their function relates more to mental health and one's status as a moral agent. Whereas Wegner's model, like Pacherie's model, concerns the generation of conscious feelings, we are concerned instead with the nonconscious processes preceding feelings such as pain and their relationships.

Two points of confusion may arise from our idea that the feeling to act is not causative and is instead explanatory of behaviour. First, how can an explanation occur before the act it seems to explain? It should be remembered here that we are not referring to *post hoc* or postdictive cognitive explanations but instead to implicit explanations or feelings. On our view, feelings represent an inference to the best explanation about the cause of the behaviour and as such it should feel as though the cause comes before the act. This is possible, we submit, if the feeling represents the predicted motor command to act rather than of the execution of the act itself. As the motor command precedes the act, so too does its prediction of cause (*i.e.* explanation). Electromyography reveals that the first muscle contractions precede the motor act by ∼90 ms (Fried et al. 2011), and as it takes ∼20 ms for signals to travel from the cortex to muscles (Robinson et al. 1988), the desire to act must occur at least ∼110 ms before the act. This short interval is consistent with our everyday experience that the intention to execute a movement occurs almost simultaneously with the movement. Second, why is a predictive explanation of the cause needed if a postdictive explanation would suffice? One answer to this question is that the feeling to act is an automatic explanation or heuristic that reduces the computational load on the brain. Rather than waste time and energy on postdictive explanations (as in the case of patient P.S.), the feeling to act is a predictive and immediate explanation for why the system acts in the way it does. However, this implicit explanation can also be used in postdictive explanations and doing so would reduce the computational burden of that process as well.

Our framework speaks to how behaviour is executed if feelings to act are not causative. It has implications for the robust discussion around the role of preceding nonconscious processing in initiating so-called 'voluntary actions' (Fried et al. 2011, Haggard 2019, Aflalo et al. 2022, Graziano 2022). While our perspective supports the view that feelings are not causative, it does not negate the experience of human agency—the feeling or belief that one has caused an action (Haggard 2017). That is, we are not suggesting that the experience of agency is not real but just that the feeling of it is not the cause of the behaviour. In our framework, feelings explain but do not cause their associated behaviours. Consider again the cutaneous rabbit illusion, which is a powerful demonstration of the brain-making sense of its internal processing. The brain appears to be using internal models to explain to itself (*i.e.* to infer causality) that a series of spatially contiguous taps should be felt as hopping taps. The brain seeks to make sense of its internal behaviour in terms of its feelings, in effect, by deceiving itself but with the added benefit of producing a sense of integrative behaviour. Imagine a creature incapable of such explanations—a half zombie, let us say, conscious of its behaviour but lacking explanatory feelings. It finds itself writhing and grimacing but feels no pain. According to the SMS hypothesis, this creature would have no immediate explanation for why it is writhing around or how it should respond. Perhaps such a being is conceivable, but without the automatic explanation afforded by pain, this creature could easily be consumed with understanding its behaviour at the expense of other life-saving actions. We propose that feelings act as heuristics to reduce computational cost by explaining to a cognitively complex system why it is behaving in the way it does. In the absence of feelings, our half-zombie would quickly become computationally overburdened.

Benefits of the SMS hypothesis

The SMS hypothesis provides insight into a long-standing question about the aetiology of a syndrome called asymbolia for pain. Some patients with this condition report being able to feel pain but they exhibit no behavioural signs typical of pain, *i.e.* patients have no reflex responses to noxious stimuli and yet they report they are experiencing pain or something like it (Griffith and Kind 2024). These patients typically have lesions of the posterior parietal cortex (Schilder and Stengel 1931, Rubins and Friedman 1948, Stengel et al. 1955), which, when experimentally lesioned in monkeys, obliterates pain escape behaviours (Dong et al. 1996). According to our algorithm, patients with asymbolia for pain would activate the second forward model in the second layer of the inferior parietal lobule, which leads to the feeling of pain without an accompanying movement (as found for the feeling of movement without movement when the posterior parietal cortex was directly stimulated; Desmurget et al. 2009). Abnormal activity within the superior parietal lobule and intraparietal cortex that prevents initiation of movement (Desmurget et al. 2018, Fornia et al. 2022) may also contribute to the failure of asymbolia patients to respond to noxious stimuli.

There is an interesting case report of four epileptic patients who during seizures have pain behaviours without any feeling of pain (Hagiwara et al. 2020). These patients are for a brief moment like our half-zombies imagined earlier. They exhibit facial grimacing, trunk twisting, limb flexion and extension, multiple body muscle contractions, grunting, and screaming during their painless seizures. This dissociation of feeling from behaviour can be accounted for by activation of the base layer of our neural algorithm and inhibition of those higher layers that lead to the pain feeling. These surprising results have clear implications for animal studies using behaviour as a test of pain.

Consider the following thought experiment. Imagine scientists are developing a drug to inhibit pain and they use rats as their experimental model. They place the rat on a hot plate and increase the temperature until the rat jumps off. Next, they determine the threshold temperature for jumping after the rat has been administered their test drug. The drug is found to increase the threshold for jumping and the scientists conclude the drug to be a remarkable success and that it should proceed to clinical trial stage. After extensive testing, however, the drug was found to have no efficacy in human patients with chronic pain, which is by far the most prevalent form of pain in clinical populations (Vos et al. 2015). What could have gone wrong? As it turns out, this is not just a thought experiment because it is widely recognized that the failure to discriminate between nociception and pain and that this has stymied translational progress on putative pain drugs (Eisenach and Rice 2022, Sadler et al. 2022, MacDonald and Chesler 2023, Palandi et al. 2023, Taylor and Ferrari 2023, Soliman and Denk 2024). In contrast to the standard framework that assumes that the function of pain is to cause behaviour, the SMS hypothesis instead raises the possibility that the drug in this scenario has simply influenced low-level circuitry, whereas the origins of pain lie in higher-order circuits associated with feelings created from

predictions of causes generated by internal models (Fig. 2). The SMS hypothesis thus focuses a spotlight on the inappropriateness of many commonly used animal behavioural tests to distinguish between pain and nociception. It suggests instead that the way forward here is to create circuit specific drugs or to modulate neural circuit activity using direct or indirect electrical cortical stimulation (Scangos et al. 2021, Zhang et al. 2023, Walder-Christensen et al. 2024)—the neural circuitry underpinning the subjective experience rather than the associated behaviour. This will only be possible once clarity on the function of pain is obtained, and once the relevant circuitry has been empirically verified.

The SMS hypothesis thus inevitably leads us to question which non-human animals have the capacity to feel if what feelings do is represent the cause of behaviour rather than be the cause. We have presented a plausible neural architecture underpinning the generation of feelings that involves forward and inverse models in series (Fig. 2). Although this architecture is not complete, it does provide a basis for determining the likelihood of which animals can have feelings. We suggest that any animal that possesses this minimal architecture or an equivalent architecture capable of executing the neural computations of these models will have, at least, the potential for subjective experience.

Conclusion

We have here proposed that the SMS hypothesis and argued that it goes some distance toward explaining how feelings and behaviour are integrated into brain models of the world. This locates the function of feelings with the way in which they inform decision-making rather than with their needing to be themselves direct causes of behaviour or actions. If correct, the SMS hypothesis is not of merely academic interest. If the assumption that the function of feelings like pain is to cause behaviour is even potentially responsible for stymying empirical research into the neural bases and effective treatment of chronic pain and other life-limiting feelings, a critical reinvestigation of the grounds for that assumption as we have proposed here is well and truly justified.

Author contributions

B.K. contributed to analysis of data and writing the manuscript. D.B. contributed to analysis of data and writing the manuscript.

Conflict of interest

None declared.

Funding

This work was supported by an Australian Research Council Discovery Grant DP200102909.

Data availability

No new data were generated or analysed in this study.

References

Aflalo T, Zhang C, Revechkis B *et al*. Implicit mechanisms of intention. *Curr Biol* 2022;**32**:2051–60.

Amanzio M, Monteverdi S, Giordano A *et al*. Impaired awareness of movement disorders in Parkinson's disease. *Brain Cogn* 2010;**72**:337–46.

Ashe J, Georgopoulos AP. Movement parameters and neural activity in motor cortex and area 5. *Cereb Cortex* 1994;**4**:590–600.

Brown R, Lau H, LeDoux JE. Understanding the higher-order approach to consciousness. *Trends Cogn Sci* 2019;**23**:754–68.

Bruno V, Castellani N, Garbarini F *et al*. Moving without sensory feedback: online TMS over the dorsal premotor cortex impairs motor performance during ischemic nerve block. *Cereb Cortex* 2023;**33**:2315–27.

Buneo CA, Andersen RA. The posterior parietal cortex: sensorimotor interface for the planning and online control of visually guided movements. *Neuropsychologia* 2006;**44**:2594–606.

Campbell JN, LaMotte RH. Latency to detection of first pain. *Brain Res* 1983;**266**:203–8.

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

Cortese A, Kawato M. The cognitive reality monitoring network and theories of consciousness. *Neurosci Res* 2024;**201**:31–8.

Darby RR, Joutsa J, Burke MJ *et al*. Lesion network localization of free will. *Proc Natl Acad Sci* 2018;**115**:10792–7.

Dehaene S, Naccache L. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* 2001;**79**:1–37.

De Havas J, Ito S, Bestmann S *et al*. Neural dynamics of illusory tactile pulling sensations. *Iscience* 2022;**25**:105018.

Desmurget M, Reilly KT, Richard N *et al*. Movement intention after parietal cortex stimulation in humans. *Science* 2009;**324**:811–3.

Desmurget M, Richard N, Beuriat PA *et al*. Selective inhibition of volitional hand movements after stimulation of the dorsoposterior parietal cortex in humans. *Curr Biol* 2018;**28**:3303–9.

Dobzhansky T. Nothing makes sense in biology except in the light of evolution. *Amer Biol Teacher* 1973;**35**:125–9.

Dong WK, Hayashi T, Roberts VJ *et al*. Behavioral outcome of posterior parietal cortex injury in the monkey. *Pain* 1996;**64**:579–87.

Duhamel JR, Colby CL, Goldberg ME. The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* 1992;**255**:90–2.

Eisenach JC, Rice AS. Improving preclinical development of novel interventions to treat pain: insanity is doing the same thing over and over and expecting different results. *Anesthesia Analg* 2022;**135**:1128–36.

Faisal AA, Selen LP, Wolpert DM. Noise in the nervous system. *Nat Rev Neurosci* 2008;**9**:292–303.

Fornia L, Rossi M, Rabuffetti M *et al*. Motor impairment evoked by direct electrical stimulation of human parietal cortex during object manipulation. *Neuroimage* 2022;**248**:118839.

Fried I, Haggard P, He BJ *et al*. Volition and action in the human brain: processes, pathologies, and reasons. *J Neurosci* 2017;**37**:10842–7.

Fried I, Mukamel R, Kreiman G. Internally generated preactivation of single neurons in human medial frontal cortex predicts volition. *Neuron* 2011;**69**:548–62.

Fried I, Wilson CL, MacDonald KA *et al*. Electric current stimulates laughter. *Nature* 1998;**391**:650–650.

Friston K, Kilner J, Harrison L. A free energy principle for the brain. *J Physiol Paris* 2006;**100**:70–87.

Garbarini F, Cecchetti L, Bruno V *et al*. To move or not to move? Functional role of ventral premotor cortex in motor monitoring during limb immobilization. *Cereb Cortex* 2019;**29**:273–82.

Gazzaniga MS. Principles of human brain organization derived from split-brain studies. *Neuron* 1995;**14**:217–28.

Gazzaniga MS, LeDoux JE. *The Integrated Mind*. NY: Plenum Press, 1978.

Geldard FA, Sherrick CE. The cutaneous" rabbit": a perceptual illusion. *Science* 1972;**178**:178–9.

Giustina A. Introspective knowledge by acquaintance. *Synthese* 2022;**200**:128.

Gläscher J, Daw N, Dayan P *et al*. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 2010;**66**:585–95.

Goldreich D, Tong J. Prediction, postdiction, and perceptual length contraction: a Bayesian low-speed prior captures the cutaneous rabbit and related illusions. *Front Psychol* 2013;**4**:221.

Graziano MS. Conscious intention: new data on where and how in the brain. *Curr Biol* 2022;**32**:R414–6.

Griffith T, and Kind A. Pain asymbolia is not pain. *Philos Sci* 2024 1–18.

Haggard P. Conscious intention and motor cognition. *Trends Cogn Sci* 2005;**9**:290–5.

Haggard P. Human volition: towards a neuroscience of will. *Nat Rev Neurosci* 2008;**9**:934–46.

Haggard P. Sense of agency in the human brain. *Nat Rev Neurosci* 2017;**18**:196–207.

Haggard P. The neurocognitive bases of human volition. *Annu Rev Psychol* 2019;**70**:9–28.

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.

Han S, Lerner JS, Keltner D. Feelings and consumer decision making: the appraisal-tendency framework. *J Consum Psychol* 2007;**17**:158–68.

Helmholtz HV. *Helmholtz's Treatise on Physiological Optics*, Vol. 3, Translated from the Third German Edition (1910). Southhall JPC (Ed.), New York: Optical Society of America, 1925.

Hohwy J. *The Predictive Mind*. Oxford: Oxford University Press, 2013.

Horgan T, Kriegel U. Phenomenal epistemology: what is consciousness that we may know it so well? *Philos Issues* 2007;**17**:123–44.

Huxley TH. On the hypothesis that animals are automata, and its history. *Fortnightly Rev* 1874;**22**:555–80.

Kalaska JF, Scott SH, Cisek P *et al*. Cortical control of reaching movements. *Curr Opin Neurobiol* 1997;**7**:849–59.

Kawato M, Cortese A. From internal models toward metacognitive AI. *Biol Cybern* 2021;**115**:415–30.

Kawato M, Hayakawa H, Inui T. A forward-inverse optics model of reciprocal connections between visual cortical areas. *Netw Comput Neural Syst* 1993;**4**:415.

Key B, Zalucki O and Brown DJ. A First Principles Approach to Subjective Experience. *Front. Syst. Neurosci.* 2022;**16**.

Lau H, Michel M, LeDoux JE *et al*. The mnemonic basis of subjective experience. *Nat Rev Psychol* 2022;**1**:479–88.

Lee SS, Wu MN. Neural circuit mechanisms encoding motivational states in Drosophila. *Curr Opin Neurobiol* 2020;**64**:135–42.

Levine J. Acquaintance is consciousness and consciousness is acquaintance. In: Knowles J, Raleigh T (eds.), *Acquaintance: New Essays*. Oxford: Oxford University Press, 2019, 33–48.

Lombrozo T. The structure and function of explanations. *Trends Cogn Sci* 2006;**10**:464–70.

MacDonald DI, Chesler AT. Painspotting. *Neuron* 2023;**111**:2773–4.

Malezieux M, Klein AS, Gogolla N. Neural circuits for emotion. *Annu Rev Neurosci* 2023;**46**:211–31.

Malle BF. *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. Cambridge, Massachusetts, USA: MIT Press, 2006.

Mashour GA, Roelfsema P, Changeux JP *et al*. Conscious processing and the global neuronal workspace hypothesis. *Neuron* 2020;**105**:776–98.

McBride J, Boy F, Husain M *et al*. Automatic motor activation in the executive control of action. *Front Human Neurosci* 2012;**6**:82.

McClelland T, Jorba M. Perceptual motivation for action. *Rev Philo Psychol* 2023;**14**:939–58.

Medendorp WP, Heed T. State estimation in posterior parietal cortex: distinct poles of environmental and bodily states. *Prog Neurobiol* 2019;**183**:101691.

Mulliken GH, Musallam S, Andersen RA. Forward estimation of movement state in posterior parietal cortex. *Proc Natl Acad Sci* 2008;**105**:8170–7.

Oatley K, Johnson-Laird PN. Towards a cognitive theory of emotions. *Cogn Emot* 1987;**1**:29–50.

Oatley K, Johnson-Laird PN. Basic emotions in social relationships, reasoning, and psychological illnesses. *Emotion Rev* 2011;**3**:424–33.

Pacherie E. The phenomenology of action: A conceptual framework. *Cognition* 2008;**107**:179–217.

Palandi J, Mack JM, De Araújo IL *et al*. Animal models of complex regional pain syndrome: a scoping review. *Neurosci Biobehav Rev* 2023;**152**:105324.

Pereira M, Megevand P, Tan MX *et al*. Evidence accumulation relates to perceptual consciousness and monitoring. *Nat Commun* 2021;**12**:3261.

Quartz SR. Reason, emotion and decision-making: risk and reward computation with feeling. *Trends Cogn Sci* 2009;**13**:209–15.

Riddoch G. The reflex functions of the completely divided spinal cord in man, compared with those associated with less severe lesions. *Brain* 1917;**40**:264–402.

Robinson LR, Jantra P, MacLean IC. Central motor conduction times using transcranial stimulation and F wave latencies. *Muscle Nerve* 1988;**11**:174–80.

Robinson WS. Epiphenomenalism. *Wiley Interdiscip Rev Cogn* 2010;**1**:539–47.

Rubins JL, Friedman ED. Asymbolia for pain. *Arch Neurol Psychiatry* 1948;**60**:554–73.

Russell B. Knowledge by acquaintance and knowledge by description. *Proc Aristotelian Soc* 1910;**11**:108–28.

Sadler KE, Mogil JS, Stucky CL. Innovations and advances in modelling and measuring pain in animals. *Nat Rev Neurosci* 2022;**23**:70–85.

Scangos KW, Khambhati AN, Daly PM *et al*. Closed-loop neuromodulation in an individual with treatment-resistant depression. *Nature Med* 2021;**27**:1696–700.

Schilder P, Stengel E. Asymbolia for pain. *Arch Neurol Psychiatry* 1931;**25**:598–600.

Soliman N, Denk F. Practical approaches to improving translatability and reproducibility in preclinical pain research. *Brain Behav Immun* 2024;**115**:38–42.

Soon CS, Brass M, Heinze HJ *et al*. Unconscious determinants of free decisions in the human brain. *Nat Neurosci* 2008;**11**:543–5.

Stengel E, Oldham AJ, Ehrenberg ASC. Reactions to pain in various abnormal mental states. *J Ment Sci* 1955;**101**:52–69.

Taylor NE, Ferrari L. Discovering chronic pain treatments: better animal models might help us get there. *J Clin Invest* 2023;**133**:e167814.

Thorell O, Ydrefors J, Svantesson M *et al*. Investigations into an overlooked early component of painful nociceptive withdrawal reflex responses in humans. *Front Pain Res* 2023;**3**:1112614.

Todd RM, Miskovic V, Chikazoe J *et al*. Emotional objectivity: neural representations of emotions and their interaction with cognition. *Annu Rev Psychol* 2020;**71**:25–48.

Vos T, Barber RM, Bell B *et al*. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the global burden of disease study 2013. *Lancet* 2015;**386**:743–800.

Walder-Christensen K, Abdelaal K, Klein H *et al*. Electome network factors: capturing emotional brain networks related to health and disease. *Cell Rep Methods* 2024;**4**:100691.

Wegner DM. Précis of the illusion of conscious will. *Behav Brain Sci* 2004;**27**:649–59.

Wegner DM. *The Illusion of Conscious Will*, New edn. Cambridge: MIT Press, 2017.

Wiech K, Tracey I. Pain, decisions, and actions: a motivational perspective. *Front Neurosci* 2013;**7**:37282.

Zajonc RB. Feeling and thinking: preferences need no inferences. *Am Psychol* 1980;**35**:151–75.

Zhang Q, Hu S, Talay R *et al*. A prototype closed-loop brain–machine interface for the study and treatment of pain. *Nat Biomed Eng* 2023;**7**:533–45.

Neuroscience of Consciousness, 2024, 2024(1), niae034 , DOI: https://doi.org/10.1093/nc/niae034, Rapid Communication
Received 10 March 2024; revised 12 August 2024; accepted 27 August 2024