ABSTRACT. A central goal for animal welfare science is the development of better theoretical and practical tools to identify and reduce animal suffering. One challenge for this project concerns how to determine which species are conscious. A second challenge concerns how we can reliably identify the behaviors of an organism that are driven by negatively-valenced states; that is, states which, if conscious, would be experienced as unpleasant. This paper aims to make progress towards this latter problem. I begin in Section 1 by introducing key terms of the debate and explain why identification of negatively-valenced states in non-human animals can be challenging. In Sections 2-4, I present three psychological characteristics of negatively-valenced states in humans and discuss how they relate to existing measures for assessing suffering in animals. I conclude in Section 5 with some reflections about the relation between the two projects of better understanding negative valence and assessing animal consciousness.

1. Introduction

Does it hurt lobsters when they are boiled alive? Do laboratory rats experience discomfort if kept severely underweight? Do farm animals in confined conditions experience emotional trauma? Questions such as these are important for animal welfare, and they concern instances in which we wish to know whether a given animal is undergoing suffering – whether they are having experiences that feel bad. Animal welfare science has developed a range of useful tools and protocols for assessing animal suffering in different species. Nonetheless, insofar as we wish to answer such questions with greater generality and confidence, we have reason to look for a firmer theoretical foundation for understanding negatively-valenced experiences in animals and their psychological and behavioural characteristics.

Two main obstacles stand in the way of such a foundation. The first concerns how we can know whether a given non-human organism is conscious, that is, whether there is “something it’s like” for an organism to undergo the experience in question. Without consciousness, it is questionable whether it makes sense to talk of an animal’s suffering (Shriver & Allen, 2005), even we might still be able to make judgments about their welfare in other terms (see, e.g., Dawkins, 2017). Some progress has been made in the science of non-human consciousness in recent years, but major theoretical controversies remain (see, e.g., Birch, 2020; Carruthers, 2020).

Quite apart from animal consciousness, however, there is a second obstacle in the way of a general theory of animal suffering, namely that of providing broader theoretical criteria.
to help us identify which specific psychological states feel bad to an organism. After all, even if we could say with a high degree of reliability that an organism was conscious, there would still remain the challenge of identifying which if any of its experiences felt unpleasant to it.

Prima facie, this might seem a relatively easy challenge to meet. After all, we know from our own case that things like pain, hunger, and nausea can feel bad, and we might reasonably generalise this observation to animals we independently believe to be conscious. Moreover, humans who live or work with animals often have well-developed standards for assessing animal welfare without recourse to broader theory (see Fig. 1 below).

Difficulties arise, though, when we consider that non-human animals often have quite different preferences from our own. Consider for example that rats show stress responses to brightly lit or noisy enclosures (Castelhano-Carlos & Baumans, 2009), but whether and when such conditions cause actual suffering to rats as opposed to merely being less preferable than dark and quiet enclosures is not a trivial question.

Similar problems arise when we consider organisms phylogenetically remote from us. Spiders, for example, seems to possess a broad capacity for nociception, recoiling from noxious stimuli, but also regularly engage in behaviours such as autotomy – the deliberate self-amputation of limbs (Johnson & Jakob, 1999). Based on evidence such as this, one might be tempted to wonder whether “arthropods in general might possess different neurobiological mechanisms for experiencing pain than vertebrates” (Kralj-Fišer & Gregorič, 2019).

Other problem cases arises from organisms with senses that humans lack. Thus consider the proposal to use powerful magnetic fields to reduce bycatch of sharks in commercial fishing (Rigg, Peverell, Hearndon, & Seymour, 2009). While this shows promise as a conservation technique, it raises the concern that the fields in question might not merely serve as deterrents for sharks but actually cause suffering.

---

1 A brief note on terminology. I will use the term ‘suffering’ to refer to instances in which a human or animal is undergoing conscious negatively-valenced states, that is, states of felt unpleasantness. These are experiences that feel bad or feel unpleasant to the organism that undergoes them, and include typical occurrences of sensations like pain or nausea as well as emotions such as fear or grief. I leave open the possibility that there may be unconscious negatively-valenced states (LeDoux & Pine, 2016), and where I intend to refer to specifically conscious occurrences, I will speak of negatively valenced experiences.
This is not to suggest that we are at a loss in tackling these cases. An increasing body of work by philosophers and scientists has attempted to better understand both behavioural and physiological indicators of suffering, and to develop generalisable instruments for assessing it (see, e.g., Sneddon, Elwood, Adamo, & Leach, 2014). This paper is intended as a contribution to this literature, and aims to connect foundational questions about the nature of felt unpleasantness to current proposed methods for assessing animal suffering.

The broad methodology I will adopt is to identify three of the central characteristics of felt unpleasantness as manifest in human experience that might enable us to make useful behavioural generalisations about its occurrence in non-human animals, and to relate these to methods in animal welfare science. I do not take myself to be giving any kind of constitutive account of felt unpleasantness, but I would suggest that the characteristics I enumerate are central enough to felt unpleasantness in the human case that they can help guide animal welfare science.

The remainder of the paper proceeds as follows. In Section 2, I argue that states of felt unpleasantness are pro tanto motivational states that play a key role in the guidance of behaviour. In Section 3, I suggest a further feature of such states (in contrast to reflex actions, for example) is that they can be fluidly ‘traded off’ against one another so as to allow for adaptive prioritization of different needs. In Section 4, drawing upon recent work by Barlassina & Hayward (2019), I suggest that states of felt unpleasantness have a distinctive self-reflexive character that may allow us to experimentally distinguish them from others kinds of motivational state. I conclude finally in Section 5 with some reflections about the relationship between the present project and the debate concerning non-human consciousness.

2. Felt unpleasantness as motivational

There are longstanding debates in philosophy and cognitive science concerning the nature of pleasant and unpleasant experience. One such debate concerns whether unpleasantness is a basic experiential quality in its own right, or merely reflects the attitudes we adopt towards different experiences. Another key area of discussion concerns how best to understand the mental content of pleasant and unpleasant states. I wish to bracket such questions in what follows. Instead, I will proceed from what I take to be robust (if perhaps not exceptionless) psychological generalisations about felt unpleasantness in humans with the goal of examining how they may contribute to practical assessments of animal welfare.

Given the present project’s focus on animal suffering, it may seem odd to begin with

---

2 For helpful taxonomies of philosophical views about the nature of pleasure (and by extension unpleasant experiences) see Sumner (1996) and Bramble (2013). For recent work analysing the content of pain and other unpleasant states, see, e.g., Klein (2015) and Heathwood (2007).
human experience. However, it is hard to see any other viable starting point. There is of course a risk of anthropomorphism or ‘anthropofabulation’ (Buckner, 2013) in making inferences about the general characteristics of a mental phenomenon while using humans as a model. Moreover, even within the human case, there are likely to be significant interpersonal and cross-cultural differences (Moore & Brødsgaard, 1999) in negatively-valenced experience. Nonetheless, the subjective nature of felt unpleasantness means that most of what we can take for granted about the phenomenon must come from our own experience.

With this in mind, I would suggest that an initial relatively uncontroversial characteristic of states of felt unpleasantness is that they are rarely if ever motivationally neutral, but instead provide a pro tanto motivation to avoid or ameliorate the state in question. If my ankle starts to throb painfully while I am jogging, it may incline me stop running or to take a painkiller, and if I am feeling intense anxiety, I may try to calm myself via breathing exercises or self-reflection.3

These are broad generalizations, of course, and matters are often not so simple. We sometimes voluntarily subject ourselves to unpleasant feelings, eating foul-tasting food so as not to offend a host or engaging in arduous exercises in the name of health. In making such cases intelligible, however, note that we normally specify some further overriding motivation that convinces us to endure the relevant unpleasant state. It is far harder to make sense of a case in which someone voluntarily undergoes or prolongs a state of felt unpleasantness for no reason whatsoever. There are of course pathological cases that may require careful handling, such as eating disorders or self-harm. However, even in these instances, psychological explanations of the relevant pathological behaviour typically invoke some further goal or some (perhaps compensatory) positive feeling accompanying any felt unpleasantness, such a desire for self-punishment or a feeling of control.4

Extending this to non-human animals, I would suggest that to the extent that a stimulus can serve an aversive function or negatively reinforce a behaviour, it has at least one of the characteristic features of felt unpleasantness in humans. We might be tempted to stop there, and suggest that we could infer the presence of negatively-valenced states directly from animal preferences. In other words, someone might suggest that the fact that rats prefer light chambers to dark ones and sharks are deterred by powerful magnetic fields gives us reason to think these stimuli result in negatively-valenced states which, if conscious, would involve suffering.

This move is too fast, however. Consider that as humans, our goal-directed behaviour

3 Reflecting this motivational role of unpleasant sensations, it is appealing to think their adaptive function is to allow organisms to respond to pressing bodily needs. See Denton (2005) for a developed view along these lines.

4 I recognise these cases are psychologically complex as well as ethically fraught. However, see e.g., Meltzer et al. (2000) for a review of some motivations often thought to be involved in cases of self-harm.
does not systematically involve direct responses to pleasant and unpleasant sensations, but reflects a variety of goals, desires, and preferences. A person might have a preference for a household temperature of 17°C, without thereby perceiving a temperature of 21°C as unpleasant. Something similar may be true of many states of non-human animals; while rats may have a preference for dark enclosures over bright ones, it does not follow that bright enclosures feel bad to rats. I will thus operate in what follows on the assumption that negatively-valenced states – states that feel bad when consciously experienced – are just one of many possible motivational mechanisms regulating organisms’ behaviour. Consequently, we face the task of determining the core features of these (but not other) motivational states.

A different sort of worry about associating unpleasantness too closely with motivation might come from instances such as ‘learned helplessness’ in which animals seem to be suffering yet display little or no motivation to escape their circumstances. This phenomenon was famously shown by Martin Seligman and colleagues, who administered electric shocks to different groups of dogs (Seligman, 1972). One group of dogs was able to terminate the electric shock by pressing a lever, while the other had no control over the duration of the shocks. In a subsequent learning task, the two groups of dogs were placed in a situation in which they received electric shocks, but could avoid them by leaping over jumping out of their enclosure. The group of dogs that had learned to terminate the electric shocks in the earlier task also readily learned this new escape behaviour, but the dogs that were formerly ‘helpless’ did not, instead simply lying down and whimpering while the shocks were administered.

The basic phenomenon of learned helplessness may seem at odds with the idea that unpleasantness is closely linked to motivation, insofar as it seems likely that the helpless dogs in Seligman’s original experiment were undergoing unpleasant experiences yet seemingly failed to take advantage of the opportunity to put an end to their shocks. Animals experiencing learned helplessness may indeed have actions open to them that would alleviate their prima facie unpleasant state, but they do not take them.

Focusing on the example of Seligman’s dogs, however, I would suggest that they may indeed have been motivated to escape the electric shocks, but unsure as to how to do so. In particular, note that the dogs did indeed commence escape behaviours if they were shown how to perform them – specifically, by having their legs moved by an experimenter in such a way as to remove them from the enclosure. In other words, once it was demonstrated to the dogs that there was indeed a viable approach to avoiding the shocks, they took advantage of

---

1 One might think that we have moral duties to help animals achieve their preferences, independent of whether these preferences are grounded in negatively-valenced experiences (see Dawkins, 2012, for such a view). I will remain neutral on this question, but would note that alleviating severe instances of animal suffering is typically taken to have special moral significance in its own right.
it. This suggests that learned helplessness may not involve the absence of motivation per se, but rather, a lack of appropriate behavioural knowledge on the part of the creature as to how to escape a given negative stimulus.

2. Felt unpleasantness as commensurable

As argued above, mere negative motivational force is an important characteristic of felt unpleasantness but not by itself a sufficient marker for negatively-valenced states: a creature might have an innate aversive response to a stimulus associated with predators, for example, without feeling anxiety or any other unpleasant state.

I turn now to what I will suggest is a second characteristic of felt unpleasantness in humans, namely that these states are to some degree intrasubjectively commensurable: we are usually able (to some extent) to quantify and compare unpleasant states just in virtue of how unpleasant they are (see Schroeder, 2004: 87), as well as gauging whether they are getting better or worse over time. Thus, if we are given the choice between drinking a foul-smelling medicine and undergoing a moderately painful injection, we can imaginatively reflect on which we would find the more unpleasant. This is not always easy to do, of course, even for a single type of unpleasant state; a patient considering a knee-joint replacement might wrangle as to whether the acute pain of post-surgical recovery was worth the alleviation of their chronic pain. Nonetheless, we can and frequently do understand unpleasantness as coming in different commensurable degrees, and take this into account in our decision-making.

Of course, as humans we possess reflective capacities unlikely to be found in many non-human animals, thus allowing us to, for example, engage in imaginative prospective reasoning about different unpleasant experiences. Nonetheless, insofar as states of felt unpleasantness are to contribute to the flexible control of behaviour as more than ineluctable instincts, we should expect animals to be able to prioritise alleviation of different negatively-valenced sensations relatively quickly.

There is evidence of this in the form of the paradigm known as motivational tradeoff. This is a phenomenon in which animals willingly undergo one negative state in order to avoid another even more negative one, or actively seek out some previously neutral stimulus if it will relieve a negative stimulus. Such motivational tradeoff behaviour in various forms has been demonstrated in a wide range of animals (see Dawkins, 2012: 150-75 for a wide-ranging summary). For example, rats, in a reversal of their normal preferences, will prefer a light chamber to a dark chamber in order to avoid unpleasant mechanical stimulation of an injured paw, while trout (highly social animals) will endure normally disliked electric shocks in order to avoid being isolated from other fish (Braithwaite, 2010: 104-5).

In all of these cases, we see evidence of animals engaging in behaviour they would normally be disinclined to perform in order to avoid some seemingly unpleasant stimulus (or
to obtain relief from an existing negative state). This suggests that they are undergoing or have undergone states that do not merely negatively reinforce individual behaviours, but which can also flexibly influence decision-making at the whole organism level. For such reasons, motivational tradeoff has been widely suggested as a promising signature of animal suffering (Birch, 2017; Godfrey-Smith, 2017; Sneddon et al., 2014).

Though commensurability may be a key characteristic of felt unpleasantness, and motivational tradeoff a useful window into animal motivations, I would suggest that it still does not quite allow us to pin down negatively-valenced states precisely enough. For one, note that apparent instances of motivational tradeoff may be accomplished by quite simple creatures such as nematode worms. Summarising recent research, Liz Irvine states that “[t]he nematode shows trade off behaviour between the ‘bad’ of crossing a strip of aversive copper ions to the ‘good’ of approaching an attractive odour, depending on the concentration of both the aversive and attractive stimuli” (2020). While it is possible that nematode worms do indeed have an internal system for representing negative states, it seems at least as plausible that they accomplish such tradeoffs via simpler mechanisms than those we might attribute to rats or chickens.

More fundamentally, perhaps, there is a worry that tradeoffs can operate not just among valenced sensations but any number of broader desires and preferences. To return to an earlier example, it may be the case that I would tolerate a less-preferred temperature of 21°C in order to spend time with a friend. It seems plausible that animals may engage in similar prioritisations, with the goal of mating taking prioritisation over a preference for a darkened enclosure, for example. Hence even if a stimulus is both normally aversive and can be flexibly prioritised over other needs, it is possible that this reflects a strong preference on the part of an animal for one state over another rather than unpleasantness per se.

3. Felt unpleasantness as self-reflexive

With this in mind, I would suggest a third characteristic of felt unpleasantness, one that may help us pin down the phenomenon yet more closely. I claimed earlier that a common feature among states such as pains, nausea, and anxiety is that they influence their subjects to act in ways that put an end to the state in question, for example, by relieving pressure on an injured body part. Crucially, however, note that unpleasant states do not just motivate us to put an end to some bodily or affective disturbance, but also motivate us to put an end to the unpleasant experiences themselves; as I will put it, they are motivationally self-reflexive.

This may seem a fine distinction, but to illustrate its importance, imagine you are suffering from post-surgical pain, with no direct form of pain relief available. In such a case, the unpleasantness of your experiences may reasonably motivate you to ask for a sedative or a sleeping pill. Depending on the nature of the drug in question, this may not alleviate the
underlying nociceptive stimulation giving rise to the pain itself, but instead remove your capacity to consciously experience it.

This thoroughgoing *self-reflexive* character of self-unpleasantness is arguably one of its central if not defining characteristics (see Barlassina & Hayward, 2019, for further discussion). In particular, note it distinguishes felt unpleasantness from other motivating states such as desires. Desires do not typically motivate me to put an end to the desire via any means available, but specifically by fulfilling them. Thus consider two people, Ali and Beckett, who are attending a party and have a reason to leave. Ali wishes to leave because she wants to return home to read her book. Beckett, by contrast, wishes to leave only because he is undergoing unpleasant feelings of social anxiety. Beckett’s desire to leave in this case is motivated by an unpleasant feeling, such that if there was an easy way to be rid of the causes of the desire (such as taking an anti-anxiety medication) Beckett might conceivably choose that option. By contrast, Ali is not motivated by any unpleasant feelings, and no such volitional ‘short-cut’ is liable to appeal to her.

I would suggest, then, that a key characteristic of unpleasant states is that they are self-reflexive, aiming at their own elimination as much as the fulfilment of other preferences (such as a preference for leaving a party) they may produce. There are of course cases where we would be glad to be rid of a desire, as in the case of someone who realizes their desire is unfulfillable and hence a source of mere frustration. However, such instances involve some further desire, whereas states of felt unpleasantness are states we are inclined to be rid of for their own sake. Translating this into behavioural terms, where a given preference (like leaving the party) stems solely from an unpleasant feeling, an individual’s preferences can be fully satisfied through the elimination of the feeling in question.

This final characteristic of felt unpleasantness allows us – in principle at least – to exclude cases of preferences in animals that lack negative affect. To give a simple example, imagine the case of an animal that had a strong preference to mate and was unable to do so, where we wished to know whether the frustration of this desire caused suffering to the creature in question. We might then place a powerful an-aphrodisiac drug in a water-bottle in the animal’s enclosure. Over time, the animal might learn that by drinking the water, its desire to mate ceased. If we found that the animal came to consistently chose to drink from this water bottle when in heat, this would provide some evidence that its unfulfilled desires were causing negative feelings, the alleviation of which it sought for its own right. Alternatively we might find that the animal showed no preference for drinking from the bottle, despite being deprived of mating opportunities. This would provide some tentative reason to think that the frustrated desire to mate was not a source of unpleasant emotions or sensations.

A related well-established paradigm for detecting negatively-valenced states in animals
is the self-administration of analgesics, in which animals with injuries show a preference for food or enclosures in which a local anaesthetic is available (Birch, 2017; Sneddon et al., 2014). So far the main evidence for this behaviour comes mainly from work on chickens (Danbury et al. 2000), but similar behaviour involving voluntary selection of environments containing analgesics have been demonstrated in other classes of vertebrates. Zebrafish, for example, choose a barren, brightly lit chamber containing an analgesic in preference to their normal enriched environment when injected with an irritating acid (Sneddon, 2013).

In these cases, it may seem prima facie as though animals are specifically motivated to put an end to the state in question, aligning with the self-reflexivity characteristic of negatively-valenced states. Some caution is still probably in order, however; we may yet find that voluntary self-administration of analgesics can be explained in terms of lower-level learning mechanisms involved in the modulation of nociception (see again Irvine, 2020, for some simple analogues in nematode worms).

The self-reflexive characteristic of felt unpleasantness might however suggest a more demanding experimental tool for assessing felt unpleasantness, namely the self-administration of general as opposed to local anaesthetics. Imagine that an animal is kept in an environment in which it undergoes prolonged exposure to some aversive stimulus such as an electric shock at regular intervals. It could be highly revealing to assess whether a creature in one such interval would ever voluntarily choose temporary drug-induced unconsciousness in preference to enduring the stimulus in question, as many humans surely would. While a negative answer to this question might not tell us much – a creature could have powerful instinctive motivations to stay awake in the presence of a perceived threat, for example – a positive answer would provide powerful evidence for the experience of negatively-valenced states. Specifically, it would suggest that a creature was motivated not simply to act in ways that reduced local nociceptive sensation, as in the case of local analgesia, but preferred temporary unconsciousness while the sensation was occurring.

5. Felt unpleasantness and the search for animal consciousness

Animal welfare science has made great strides in recent decades, and we are considerably closer to understanding how negative stimuli are processed by many animals. Nonetheless, insofar as we wish to move towards more powerful accounts of animal suffering that can generalise to creatures with radically different preferences, nervous systems, or sense organs, we have grounds for seeking to better understand the phenomenon of felt unpleasantness itself. In this paper, I have outlined one pathway to making progress in this regard, namely by seeking to identify general characteristics of felt unpleasantness in humans and extrapolating from these to hone and innovate our empirical techniques for assessing animal suffering. While I would endorse adoption of the precautionary principle (Birch, 2017) in making
inferences about the absence of negatively-valenced states in a given species, the broad approach outlined here could at least provide powerful positive evidence for their occurrence.

Throughout this paper, I have largely (and deliberately) bracketed questions of animal consciousness, but as noted at the outset, it is questionable whether non-conscious instances of negatively-valenced states have any direct bearing on animal welfare. Thus even if we had clear evidence that a given organism underwent negatively-valenced states of the same broad psychological kind as those involved in human experiences of felt unpleasantness, if these states are unconscious this may not amount to an instance of suffering. Nonetheless, it can be hoped that progress towards identifying and understanding negative-valence across species can advance in parallel with insights from the science of non-human consciousness.
REFERENCES


