Cognitive bias in animal behavior science: A philosophical perspective

Behzad Nematipour*, Marko Bračić, Ulrich Krohs

Affiliations

BN: Center for Philosophy of Science, University of Münster, Domplatz 23, 48143 Münster, Germany
MB: Department of Behavioural Biology, University of Münster, Badestr. 13, 48149 Münster, Germany
UK: Department of Philosophy, University of Münster, Domplatz 23, 48143 Münster, Germany

Correspondence

Behzad Nematipour, behzad.nematipour@uni-muenster.de

ORCiD
BN: 0000-0003-0329-4107
MB: 0000-0001-6528-3572
UK: 0000-0002-5524-9626

Abstract

Emotional states of animals influence their cognitive processes as well as their behavior. Assessing emotional states is important for animal welfare science as well as for many fields of neuroscience, behavior science, and biomedicine. This can be done in different ways, e.g., through assessing animals' physiological states or interpreting their behaviors. This paper focuses on the so-called *cognitive judgment bias* test, which has gained special attention in the last two decades and has become a highly important tool for measuring emotional states in non-human animals. However, less attention has been given to the epistemology of the cognitive judgment bias test and to disentangling the relevance of different steps in the underlying cognitive mechanisms. This paper sheds some light on both the epistemology of the methods and the architecture of the underlying cognitive abilities of the tested animals. Based on this reconstruction, we propose a scheme for classifying and assessing different cognitive abilities involved in cognitive judgment bias tests.

Keywords: ambiguous stimuli; cognitive bias; judgment bias; emotions; representation; decision-making

Statements and Declarations

This paper was funded by the German Research Foundation (DFG) as part of the SFB TRR 212 (NC³) – Project number 316099922. There is no competing interests or funding with regard to this paper

Acknowledgements

We would like to thank Helene Richter, Norbert Sachser, Rose Trappes and Peter Schulte for comments on previous drafts. Thanks are also due to all members of the CRC TRR-212 at the University of Münster, Bielefeld University and Jena University and to the participants of the colloquium in the theoretical philosophy from SS20 to WS20/21 at Münster University for their constructive feedback on the project.

1 Introduction

Assessing animals' emotional states has explanatory, predictive, and illustrative value for animal welfare science, neuroscience, and psychopharmacology (Mendl et al. 2009), as well as for the discourse concerning attributing rights to *sentient species*. However, this assessment is particularly difficult in non-human animals because of the lack of verbal interaction. That is why scientists in these fields are looking for various indicators of emotional states such as behavioral and physiological changes that accompany such states in order to assess in which emotional state an animal is, or whether or not animals of the considered species have them at all (Kremer et al. 2020). For example, the state of fear may be accompanied by behavior like freezing, fleeing, or even attacking, and by physiological changes such as increased heart rate, increased blood pressure, and enhanced levels of circulating glucocorticoids (Mendl et al. 2009). In biomedical research, animal models for emotional disorders, such as anxiety and depression, are often based on exposing animals to stressful conditions and then recording behavioral indicators, e.g., immobility, exploration versus avoidance, self-grooming, and vocalizations (see Bourin 2015; Kalueff et al. 2016; Simola and Granon 2019; Wang et al. 2017). Animal research in general uses emotional indicators mostly to detect negative emotional states (Proctor et al. 2013) and the methods for assessing positive states are limited (Paul et al. 2005). This limits the research of emotions in non-human animals, particularly from the perspective of animal welfare because of the aspiration to induce positive states, in addition to reducing pain and suffering, in animals (Boissy et al. 2007). Additionally, many commonly used indicators suffer from certain epistemic problems (which are discussed in detail in the next section). This motivates scientists to consider novel indicators that are potentially more reliable and can also detect positive states (Kremer et al. 2020).

An increasingly used indicator of emotional states in non-human animals is *cognitive bias* (Paul et al. 2005). This indicator has its background in psychological experiments on humans; emotional states affect our memory, attention, and judgment (Mathews et al. 1995; Mineka and Sutton, 1992). A paradigmatic example of such influences is that people in negative emotional states, like anxiety, depression, or fear, tend to remember and focus on negative events and interpret ambiguous situations negatively.

The potential utility of testing cognitive bias in welfare research was demonstrated in the seminal study of Harding et al., who showed that rats housed in "unpredictable"/stressful conditions (which cause depression-like symptoms) were inclined to respond more negatively to ambiguous situations than rats housed in "predictable"/familiar conditions. Their judgment was biased (Harding et al. 2004).[1] The authors suggested using behavioral responses in ambiguous situations as an indicator of emotional states (Harding et al. 2004; Paul et al. 2005), which initiated numerous studies that demonstrated that cognitive judgment bias can be found in a wide range of taxa, from pigs to bumblebees (reviewed in Lagisz et al. 2020; Mendl et al. 2009; Neville et al. 2020). Since both pharmacological and environmental manipulations of affective states alter judgment bias (reviewed in Lagisz et al. 2020; Neville et al. 2020), cognitive judgment bias tests can be considered as a promising tool for assessing emotional states of non-human animals.

In this paper we pursue two main goals. First, we want to examine the epistemic role of judgment bias as an indicator of emotions. We start by pointing at known epistemic problems with the more traditional indicators of emotional states (behavioral and physiological changes) and point at some advantages as well as limits of using judgment bias as an indicator of emotional states in animals. We aim at assessing the epistemic value of the judgment bias test and demonstrate its empirical motivation.

Second, we scrutinize judgment bias as such. What kind of cognitive abilities are in play? We are not engaging, however, in a conceptual analysis of the notion of judgment bias, but rather looking at cognitive abilities underlying the judgment bias that is used as an indicator of emotional states, and aiming at determining what kind of abilities these are. Animal welfare science might not need to determine exactly what kind of ability is used as an indicator as long as there are proper ways of tracking or individuating the emotional states.

[1] What exactly the housing conditions were and what it meant by responding "more negatively to ambiguous situations" will be clarified and discussed in section 2.

From other perspectives, however, this question is worth pursuing, because (1) for cognitive science and philosophy of mind, the (exact) kind of cognitive abilities of non-human animals is germane to understanding language acquisition and the evolution of higher cognitive abilities. (2) Pinpointing underlying cognitive abilities in different species might clarify minimal requirements for cognitive and emotion-like systems to produce such a phenomenon. (3) Even from the perspective of animal welfare studies, there are disparities between treatments of animals with higher and lower cognitive abilities. Therefore, it might be important to determine which level of cognitive abilities is in play in cases of judgment bias. This is important because evidence of judgment bias across the animal kingdom has fueled a heated debate on attributing emotional states and consciousness to species that are usually not considered being sentient (Mendl and Paul 2016) – a debate that has ramifications for questions concerning animal welfare and animal rights.

2 Epistemic limits of emotional indicators

Most scientists seem to agree that at least some emotional states can be ascribed to (some) non-human animals (Scarantino and Sousa 2021). However, in affective science there are not only many different theories on what constitutes emotions, but also the terminology is inconsistent, which can create misunderstandings when discussing emotions in non-human animals (see e.g., Adolphs and Andler 2018; Barrett et al. 2007; Izard 2010; LeDoux 2012; Paul and Mendl 2018). Therefore, before discussing the limitations of emotional indicators, we need to clarify the terminology. In this paper, we use the term “emotional state” broadly, referring to inner representational states without presupposing subjective or conscious experience. Concerning the structure of emotional states, we try to generalize across both, discrete approaches – considering basic emotions as discrete entities, underlayed by separate neurological systems (see Ekman 1992) – and dimensional approaches – specifying emotions by the position in multidimensional space, with two common dimensions being valance (i.e., pleasantness or unpleasantness of emotional state) and arousal (i.e., activity or energy level) (see Russell and Barrett 1999).

Since emotions are not directly observable, assessing emotional states requires the use of indicators. There are two major types of problems with emotional indicators like behavioral and physiological changes. First, they are not unique to a specific emotional state. In other words, two or more different emotional states could be accompanied by the same/similar physiological and behavioral changes. This means that the indicators are not always indicators of *uniquely one* emotional state (e.g., fear) or emotional dimension (e.g., unpleasant). This is problematic in biomedical and animal welfare research when trying to assess whether an animal is in a specific emotional state. For example, detection of an elevated level of circulating glucocorticoids as compared to the baseline could indicate that the animal is in a negative state (e.g., fear), but the same effect would be observed if the animal is aroused positively and thus, in a positive state (e.g., reward anticipation). Without appropriate context, the elevated glucocorticoid level thus turns out to be an indicator for emotional arousal in general rather than indicating a negative state (Ralph and Tilbrook 2016). Play behavior, to take another example, is generally considered a good indicator of positive emotional states, but in some cases, increased playing activity was connected with a negative emotional state of the tested animal as assessed by an independent method (Ahloy-Dallaire et al. 2018; Richter et al. 2016). Consequently, even commonly used indicators can fail to indicate the emotional state correctly when considered alone or taken out of biological context. Let us call this type of problem *the specificity problem* of emotional indicators.

The second type of problem concerns the reliability of the emotional indicators *as indicators*. The observed physiological and behavioral changes are not exclusively caused by emotional states. A specific change in the behavior or physiological state of an animal could be caused by, for example, an adaptive coping mechanism that does not involve any emotional states. Moreover, certain stereotypic behaviors are unreliable indicators of emotional state. It has been shown that they could be a direct way to cope with a stressor (e.g., poor housing conditions) directly by "do-it-yourself-enriching" the environment or by calming themselves and thus, blocking stress, rather than coping with the condition indirectly *via* first eliciting another emotional state that then lowers stress (Mason and Latham 2004). Let us call this type of problem *the reliability problem* of emotional indicators.

One way to overcome these epistemic difficulties is to look for new ways of assessing animal emotional states that are: (1) more emotion-specific; (2) more reliable; or (3) give rise to more reliable and/or emotion-specific indicators *in combination with* already existing indicators. A relatively new and popular emotional indicator that is assumed to overcome these problems is judgment bias. Before scrutinizing the explanatory power of judgment bias experiments, let us see how exactly these experiments are set up.

Cognitive judgment bias test. Judgment bias experiments were, among other things, designed to show that decision-making and judgment in non-human animals are influenced by emotional states. The design of a judgment bias test can be generalized as follows. Animals are first trained to respond differently to two distinct cues of the same perceptual dimension (e.g., visual, auditory, or spatial cues): they learn to respond to a "positive" training cue to obtain a reward (e.g., lever press for a high-pitched tone), and to respond in another way to a "negative" training cue to avert punishment (e.g., no lever press for a low-pitched tone).[2] This is the *training phase*. When the animals have learned to respond correctly to training cues, they proceed to the test where they are presented with "ambiguous" cues, which are interspersed between training cues – this is the *testing phase*. Ambiguous cues are located qualitatively between the training cues associated with negative and positive effects – hence ambiguous – and are usually not rewarded. The behavioral response to ambiguous cues is considered to indicate whether the animal "anticipates" positive outcomes (responding as expecting reward) or negative outcomes (responding as avoiding punishment). These responses are shown to be sensitive to a change in emotional state (Lagisz et al. 2020; Neville et al. 2020) and they could be used as indicators of emotional states.

For interpreting the observational data, it is required to know in which emotional state the animals are when responding to the ambiguous cues. Therefore, they are manipulated in a (emotionally) *priming phase* to be in a certain emotional state before being tested (in the testing phase). A typical setting divides animals into two groups. One group is manipulated by a treatment considered to induce a negative emotional state. The other serves as the control group and stays unmanipulated. Priming could be an unpredictable housing, lasting throughout the training and testing phase of the experiment, or an enforced electrical shock applied just before the testing phase. Similarly, positive treatment can be used as priming (e.g., providing enrichment).

We could observe the following outcomes of such a judgment bias experiment:

1. The animals from the "negatively" primed group might show a negative judgment bias by responding more often in the negative way with respect to the ambiguous cues, i.e., by the behavior they have learned to avoid punishment, than animals from the control group. The interpretation of this result would be that animals from the "negatively" primed group are in a negative emotional state and thus, that this particular priming is inducing a negative emotional state. The same goes for positive treatment.
2. Priming might also not lead to any change in the interpretation of the ambiguous cue, so that there is no statistically significant difference between the treatment and the control group. This might mean that the priming did not evoke any emotional state that lasts until

[2] The association of a specific cue (e.g., tone pitch) with reward or punishment is usually counterbalanced. Also, there are test designs based on so-called go/go tasks where animals always need to respond to both "positive" and "negative" cues, which removes the confounding effect of motivation to respond.

the judgment bias test, or that the bias is too small to be detectable with a concrete experimental setting. However, as long as data from different individuals are averaged, it might also be the case that individuals react to the same treatment with different emotions, some with positive ones and others with negative ones.

Let us now come back to the epistemic problems with emotional indicators. It seems quite obvious that the judgment bias test inherits the reliability problem. The described experiments may show that there is a correlation between a treatment, expected to induce certain emotional states, and behavioral responses in a judgment bias test (as described in outcome 1). However, different factors, besides the emotional state, can influence judgment bias (Whittaker and Barker 2020). Consequently, if it is possible and plausible that cases of judgment bias could occur without any involvement of emotional states, then judgment bias has the same reliability problem as other indicators. This, of course, does not mean that the overall reliability could not be increased if we took additional indicators into account. The point is rather that if we look at each emotional indicator (including judgment bias) separately and try to assess the emotional states by it, then the reliability problem remains unsolved.

At first glance, it seems that the specificity problem, too, accompanies judgment bias as an emotional indicator. It is hard to imagine that one could be able to specify the exact kind of emotional state of an animal (fear, anger, depression-like states, joy, frustration, etc.) just from the judgment bias, be it negative or positive. However, the experiments seem to suggest that there are correlations between negative bias and negative emotional states in general and between positive bias and positive emotional states in general (Lagisz et al. 2020). This is certainly a relevant differentiation and might, in many cases, even be sufficient from the particular perspective of animal welfare scientists, because, as mentioned before, their interest often is assessing whether or not animals are in negative (or positive) emotional states. So while, for example, an elevated level of circulating glucocorticoids could be indicating either a state of fear or one of excitement and thus, does not allow to infer a negative or a positive emotional state, a state of fear would usually correlate with a negative judgment bias and a state of excitement with a positive one. In this respect, judgment bias promises to be more specific than some other indicators.

To sum up, in light of inherent epistemic problems of traditional emotional indicators, there are at least two reasons to consider judgment bias tests as an alternative: (1) Where emotional indicators have different degrees of reliability, having another indicator *in addition to* the already existing ones can increase the overall reliability of these indicators when all are pointing to the same emotional state; (2) With a cognitive bias test, it seems possible to assess whether a certain treatment *induces* a positive or negative emotional state, which is of eminent value for animal welfare science and biomedical science.

Having discussed some inherent problems with emotional indicators and established the epistemic motivation of judgment bias tests, let us now discuss the underlying cognitive mechanism.

3 Underlying cognitive abilities

We are now going to scrutinize cognitive abilities that could underline and explain behavioral responses to ambiguous situations in judgment bias tests. However, before introducing possible candidates, let us first specify the category of cognitive abilities. With this category we are referring to mental capacities, like the abilities to represent, to have emotions, to perceive, to judge, and other higher level mental capacities. These are to be discerned from the neuronal activities and processes that form the basis of these capacities. Such more fundamental processes cannot *as such* explain animal activities and behaviors in judgment bias tests; the concept of judgment bias focuses on representational states rather than on their neuronal basis. This can be seen in both "folk psychology" and empirical sciences. Consider, for example, answers to questions like: "Why is that squirrel climbing that tree so fast? Why is that honeybee flying in that direction?" The answers would usually refer to representational states or abilities rather than to – unknown – neural states: because the squirrel is *scared of* and *running away from* the dog (representing it *as dangerous*) or because the honeybee *represents* the nectar to be in that direction, say, as a result of *observing* the dance of a fellow bee and *interpreting* it *as representing* the nectar occurrence in that direction. Analogically, the answer to the question of why the animals in the cognitive bias tests respond to the ambiguous cue in a specific way would refer to some kinds of emotional or inner representational states or ability. That is why ascertaining possible representational abilities that can result in the behaviors in question has immense explanatory value for scientists conducting cognitive bias tests.

A last remark before our analysis of the underlying mechanisms; this is not an analysis of the terms "ambiguous," "bias," or "judgment" or of their applications. Our listed candidates of inner states and abilities that explain the reaction to the ambiguous cue in the judgment bias tests may or may not confirm the usage of these terms – whatever the criteria of this confirmation might be. Nevertheless, our focus is not on this kind of conformation but on plausible candidates for different cognitive abilities that would result in similar behavioral outputs with similar input conditions in these tests.

3.1 Plausible candidates

Scientists experimenting on judgment bias often do not ask the question about the (exact) kind of cognitive abilities that bring the bias about. They are very cautious in classifying the responses as merely being “as if” the animal expected a certain outcome (Mendl et al. 2009; Paul et al. 2005; Roelofs et al. 2016). Usually, they treat the involved cognitive mechanism as a black box and track it through its behavioral outputs.[3] As clarified before, we think that this question is worth answering from both perspectives, that of cognitive science and that of animal welfare science, as it could lead to refined measurements, development of new tests, and better understanding of emotional states in general.

Our approach to answer the above question is to make a list of cognitive abilities that are discussed in philosophy of cognition and that we, at the same time, consider as being evolutionary plausible candidates that might produce the biased output in a systematic or regular way. This will outline some of the possible and plausible underlying abilities that contribute to the mechanism in the assumed black box. The answer would in part require describing some inner states of the animals *as* representing the external cues, i.e., assuming that the states are *directed at*, or *are about* an external phenomenon or state of affairs (e.g., Sterelny 1990).

1. Constitutive lack of discrimination. It is plausible that the cognitive system of some animals does not discriminate between the cue that, *from our perspective*, should be ambiguous for them, and one of the training cues. This inability might be a “constitutive” lack of discrimination between ambiguous and training cue, and would not be mediated or altered by emotional states and other conditions, for it is a matter of physiology and unmodifiable by priming. Imagine, for example, somebody who suffers from a particular kind of color blindness and cannot discriminate between, say, blue and purple but can distinguish red. This person now receives a purple cue, meant by the experimenter as a middle cue between blue and red and, and sees it as blue. The test person’s perceptual apparatus simply does not discriminate between what we would classify as a middle cue and as one of the others. Now

[3] Mendl et al. (2009) sketched a picture of what they hypothesized as underlying mechanisms of judgment bias which we will in part discuss in this section. However, they admitted that this might not concern animal welfare studies in practice: “From a practical animal welfare perspective it is perhaps not necessary to understand the processes underlying judgment biases” (ibid. 172).

imagine that this is the “normal” case for the whole species that is being experimented on; the cue would not be ambiguous for individuals belonging to this species.

This possibility is eliminated if animals show the ability to discriminate between the cues in a separate experiment or if animals respond differently to ambiguous cues than to the training cues in the judgment bias test. This seems to be the case in most published studies since different responses to at least some ambiguous cues are considered a prerequisite for a valid judgment bias test (Gygax 2014, 61). We are mentioning this case for reasons of completeness, and also because it helps to better understand the other candidates.

2. Misrepresentation. One of the most plausible situations that might hold is that the ambiguous cue is represented – wrongly – as one of the cues the animal was trained upon, i.e., that it is misrepresented (Dretske 1986; Godfrey-Smith 1989). Assume the cues trained upon were squares and circles, and the ambiguous cue being an octagon. If the content of the representation is an octagon (however, one could possibly find this out), the ambiguous cue would be represented correctly. If the ambiguous cue is represented either as a circle or as a square, or in the very way a circle or a square is represented, it is misrepresented. Or consider the following standard example of a misrepresentation (Agar 1993): a frog *misrepresents* a certain black particle, let us say a small black piece of paper, in the air as, say, a nutritious flying prey, and the prey-capture mechanism of the frog triggers a tongue-dart in the appropriate direction and captures the piece of paper. This could happen for various reasons; the black piece of paper looks just too much like a fly or the frog is just too hungry etc. The point is that the piece of paper is not represented as a piece of paper (which would be impossible as long as we assume that this category does not exist at all for the frog), and that it is also not the case that it is not represented at all. It is represented as something else with which the frog is familiar, in this case as a fly. It is likely that something similar is happening in judgment bias experiments when an animal observes an ambiguous cue; the cue is misrepresented as a “familiar one.”

Mendl et al. mention that something like this might be the case with ambiguous cues that are very similar to training cues but argue that this is likely not to be the case when ambiguous cues can be easily distinguished from training cues (Mendl et al. 2009, 172). So, to exclude misrepresentation, does one simply need to confirm that animals can discriminate between ambiguous and training cues in classical discrimination experiments where two cues are presented simultaneously? This would be too quick a conclusion. Consider the following: just because one is able to distinguish between cats and dogs under ideal or standard conditions, it does not mean that one is not likely to confuse them under certain circumstances or in certain contexts, e.g., to mistake in dim light a small dog for a cat. Similarly, just because animals showed the ability to discriminate between the ambiguous cues and the training cues, they need not be able to do so under testing conditions of judgment bias experiments, where multiple ambiguous and training cues are presented sequentially with time gaps in-between (as mentioned by Mendl et al. 2009, 173). They might still misrepresent ambiguous cues as one of the training cues. Misrepresentation can occur for various reasons. The reward is just too delicious, or at least delicious enough to mistake anything *resembling* the positive cue as *being* the positive cue; or the punishment is too severe or severe enough so that anything resembling the negative cue gets mistaken as being the negative cue; or the emotional inducing phase made the test animals too cautious, too afraid, too anxious, too bored etc.

As an argument for a more advanced cognitive ability, Mendl et al. use the observations that there is a gradual change in response to cues in judgment bias tests (Mendl et al. 2009, 173). In a typical judgment bias test, animals are often introduced to three ambiguous cues; one ambiguous cue is closer to the positive (near-positive), one is closer to the negative training cue (near-negative), and one is perceptually in the middle. This scheme is applied to test whether there is a gradual change in animals' responses across the cues. For example, animals reduce lever pressing from the positive cue *via* ambiguous cues to the negative cue, thus producing a monotonic response curve. If there is a gradual change in responses, it is presumed as validating that animals interpret ambiguous cues in reference to the trained cues (e.g., Gygax 2014, 61; Hintze et al. 2018, 10). Assuming that the middle ambiguous cue is not perceived as actually being one or the other of the training cues, Mendl et al. consider that it is likely that something cognitively more advanced like decision-making is happening. Although we grant that something like this might be happening in animals with higher cognitive abilities, which we will consider next, we want to emphasize misrepresentation as being one of the most likely scenarios, even in cases where one might consider decision-making as being an alternative mechanism. To be clear, our estimation of likelihood here is not based on empirical data but rather on the principle of Ockham's razor to be as scarce as possible with assuming entities, in this case with presupposing involved cognitive instances or abilities. In fact, if it is likely that the animals are misrepresenting the ambiguous near-positive and near-negative cues, we do not need to – and should not – bring some higher cognitive abilities, like decision-making, into play to explain the response to the middle ambiguous cue as long as there is no concrete indication for involvement of the higher capacity. In the case mentioned, there is no such independent argument for the presumption that the animals' gradual responses indicate decision-making.

Nevertheless, because it is possible that more advanced cognitive abilities would produce the similar output under the similar input conditions (as in the case of humans), we will still consider this option and try to identify the minimal requirements of such a cognitive system according to an evolutionary perspective.

3. Conflicting content(s). The third possibility which could be available in an advanced cognitive system is the representation of the ambiguous cues *as* ambiguous, for example, as something undetermined between two or more *specific* states or objects. To have an analogy from the perspective of a (human) viewer, it is *not* like: "I am seeing something but I don't have any idea what it is," but more like "I am seeing something that is either *x* or *y*, but I cannot exactly tell which one of those two." The latter is analogous to the cases that we are considering now.

It is important to note that the conflicting content(s) could be different contents of different representations of the same state of affairs, or a "conflicting" content of a single representation of that state of affairs. Without going too deep into the theories of content, with a *conflicting* content of one representation we are referring to a content that has two or more aspects with different psychological roles (hence "conflicting"), e.g., a state of affairs is represented as being a dog or a cat or even as a dog or a non-dog, where there are different behaviors associated with these different aspects, for example fleeing in case of the representation of a dog and attacking in case of a cat or a non-dog. How exactly these aspects are represented and how the connections between them appear is not relevant here. It is merely relevant that the cognitive system links these different aspects to different behavioral outputs[4] and that the cognitive system has the means to deal with this conflict.

While it might sound natural that humans have such representations, the issue is much more complex than it appears at first glance. In general, the state of affairs in question needs to be represented *as* conflicting (either through the conflicting representations or the conflicting aspects of a representation of the state of affairs), which furthermore means that there are mechanisms, over and above "regular" representational mechanisms, that evaluate these representations and compute, or "decide" about,[5] the generation of an output signal that enters the behavior-producing mechanisms. This feat of the cognitive system is a capacity

[4] "Behavioral output" is to be understood in a broad sense and does not need to be a behavior of the animal in the strict sense. It includes, for instance, activities of some subsystems that are triggered by the representation(s).
[5] We assume "computing/'decision-making'" as not necessarily being a conscious process.

over and above the ability to represent (and misrepresent) something in a specific way, because simple representational systems do not usually evaluate representations or aspects of a representation *against each other*.

We want to emphasize that we are not suggesting that there is no evaluation of representations or some kind of computing happening in cases of mere misrepresentations. However, if the animal has a representation with a conflicting content or competing representations, then some kind of *resolving*-mechanisms should come into play that deal with the ambiguity. Doing this in a consistent way requires the involvement of a different, more advanced cognitive ability than would be required in reacting to a mere misrepresentation of a cue. Bear in mind that from the setup of the judgment bias experiments there is not yet much known that allows us to assess which kind of these cognitive abilities (misrepresentation *versus* conflicting contents) is in play. Our analysis suggests a way of gaining better knowledge about the representational systems, i.e., a way to open the black box at least a little bit; does the animal always react the same way to an ambiguous cue, or does it learn to distinguish it from the training cues? One might expect that conflicting content is interpreted cautiously or with hesitation on the first confrontation, but more decisively in later ones, while a plain misrepresentation would not give rise to any hesitation.

A judgment bias test, however, would merely hint at certain mechanisms and cannot be used to conclusively distinguish between cases of misrepresentation and of conflicting contents. We will therefore discuss, in section 3.2, more complex experimental setups that could yield more definite results on the representation mechanism involved.

4. Novel representation. The last option that we want to consider is the possibility of having a *novel* representation, i.e., to represent the state of affairs – the ambiguous cue – *as novel*. To use the analogy from before, it is more like: “I am seeing something but I don’t know what exactly it is.”

Representing something as novel does not mean that the representation is marked by a “novel”-index. It also does not mean that the one having the representation “thinks” the content is novel (conscious or not). All it means in this context is that the one having the representation has not yet gathered any prior information about what is being represented, which includes in particular that it does not relate the novel representation as being related to the training cues. One of the most central feats of the cognitive system is to use information gathered in prior encounters with an entity in the current or future encounters with that entity (Millikan 2000). It is therefore common for the cognitive system to start tracking and gathering information about newly encountered unknown entities.

Representing ambiguous cues as novel is more likely in certain types of judgment bias tests. The most prominent case is when the cues do not differ in only one perceptual dimension (e.g., Douglas et al. 2012; Nogueira et al. 2015; Salmeto et al. 2011). For example, Douglas et al. (2012) used different acoustic sounds: a note on a glockenspiel and a dog-training clicker as training cues, and a squeak from a dog toy as the cue which was considered to be ambiguous. However, do animals perceive these sounds to be different in frequency, in noise level, or in some other dimension? In such cases, it is not clear how animals relate ambiguous cues to training cues; they could be represented as novel (as mentioned by Roelofs et al. 2016). Although for a different reason, novelty could also play a role in judgment bias tests that are based on spatial cues. In this type of tests, ambiguous cues are represented by a novel location which is in-between the trained cues (e.g., Briefer and McElligott 2013; Richter et al. 2012). Jardim et al. showed that in this design, reaction to the ambiguous situation depends on how explorative an individual is and thus, includes the animal's response to novelty (Jardim et al. 2021). It is thus possible that animals represent situations that are intended to be ambiguous as novel, at least in some judgment bias tests.

The behavioral output in such cases would depend on various factors, such as the level of individual development of the cognitive system, the individual's prior learning experiences, the overall cognitive capacities of the species, the organism's predispositions, and of course the organism's present environment and emotional state. However, if the organism had a genuinely novel representation, it could be expected that it would change its behavior depending on the kinds of information being gathered about the entities in question (here, the ambiguous cue). For example, if the ambiguous cues are not associated with any reward or punishment and the animal starts to ignore these cues pretty quickly, it would suggest that (at some point) the animal has had a novel representation of the ambiguous cue and that the representation has a different content than the representations of the previously learned ones. This so-called "loss of ambiguity" is observed in many studies (reviewed in Roelofs et al. 2016).

Before describing our suggestion about (practical) ways of differentiating these options, let us make some important clarifications. Firstly, we do not suggest that our list of possible candidates for mechanisms is complete. This is the list of options that we think are the most plausible candidates for the underlying cognitive abilities. Others might be possible. Secondly and most importantly, we do not think that these possibilities are mutually exclusive. In other words, it is possible that the underlying cognitive ability of a process studied is a complex combination of these options. For example, an animal could misrepresent the ambiguous cue at first but start perceiving it as novel later and change/adjust its behavior accordingly; or the animal could perceive the cue as novel but misrepresent some aspects of it as being dangerous or advantageous and so on. This means that assessing the exact configuration of the underlying cognitive mechanisms through experiments requires thorough planning, more complex training phases (we will address this in the next section), and various controlling scenarios, which taken together might be near impossible to conduct for some species. Nevertheless, in the next section, we will suggest a setting that is less likely to involve *misrepresentation*.

3.2 Ways of differentiating: A new proposal

As we stated earlier, the possibility that the tested animals might lack the ability to discriminate between the ambiguous cues and the cues in the training phase can be eliminated through separate experiments that test their perceptual abilities. However, things are more complicated if we are to establish whether a behavioral output of the judgment bias test is the result of a *misrepresentation*, of *conflicting contents*, or an instance of *novel representation*.

As a promising way that is less likely to involve misrepresentation than *conflicting contents* or *novel representations*, we propose using a setting that involves two pairs of cues during both training and testing phases.[6] In the training phase, animals need to learn associating two different cues[7] with negative and two with positive outcomes. In the testing phase, rather than using novel cues that are supposed to be ambiguous, the properly conditioned animals are exposed only to cues they are already familiar with, namely to a combination of one cue that is associated with a negative outcome and simultaneously to another one associated with a positive outcome. In this kind of experiment, a conflicting input is realized by combining the positive cue of one of the pairs with the negative cue of the other at the same time. Therefore, in contrast to a judgment bias test, the "ambiguity" is represented not by one ambiguous cue, but rather by two different, conflicting cues. Each of the cues is unambiguous and might even address different sensory modes (e.g., visual and auditory).[8] It is important to test both options of "ambiguous combinations" of cues, a positive cue 1 with a negative cue 2 and a negative cue 1 with a positive cue 2. This rules out that one of the cues might generally override the other. Individuals would need to provide relatively consistent

[6] This setting has been used in Parker (2008) for a different purpose than we are proposing here.

[7] Optimally, both senses should have similar perceptual values for the animals to avoid the possibility that the behavioral outcome is the result of the animals being overly sensitive to one cue rather than the other.

[8] Of course there should be several control groups with negative-negative, positive-positive, and negative-positive with a different timely distance between the sensorially different cues.

answers to both "ambiguous combinations" for the experiment to be valid, i.e., ascribing ability to solve conflicting content. Completely random answers of the individual would suggest a lack of relevant problem solving mechanisms.

This experiment is not supposed to be an improvement upon the judgment bias test. The suggested setup serves the purpose of singling out specific mechanisms underlying the judgment bias, in the case of scientific interest in doing so. This setup is primarily supposed to test whether animals possess certain conflict-solving mechanisms. However, it does not necessarily exclude a novel representation of the presented conflict.

Possible outcomes. In the following, we discuss the possible outcomes of such an experiment and show which conclusions could be drawn with respect to how the underlying cognitive system represents the cues:

The individuals are trained to the cues and then exposed to ambiguous combinations of cues, without any prior exposition to emotion-eliciting conditions. Let us assume that the punishment and rewards in the experiment are "fair," i.e., they are not too highly evaluated by the animals.[9]

A. Each individual might show a consistently biased answer, positively in some individuals and negatively in others. This would allow ascription of a dispositional trait to the individuals that count as long-lasting. We could call these individuals "optimistic" or "pessimistic" decision-makers.
B. All individuals might show a similar bias, either positive or negative. One could interpret this as constitutive optimism or pessimism being a certain dispositional trait of the species under investigation, where either a positive or a negative cue overrides an opposing cue. Existence of such "optimistic" or "pessimistic" species traits might be expected if they were selected for due to certain living conditions.[10]
C. The answer might be found to be arbitrary in all individuals, i.e., the ambiguous combination of cues leads to positive and negative answers in statistically indiscernible proportions in each individual. The conflicting contents, which in isolation lead to a positive and negative answer, respectively, level out. This outcome would strongly suggest that the animals do not possess the relevant problem solving mechanisms at the

[9] Finding out whether or not the reward and punishment are evaluated "fairly" by the animals would involve prior experience and experiments which might differ from species to species.

[10] The same outcome would be expectable if the punishment or reward are evaluated too highly by the animals. However, this option should be excluded by proper test design, so the explanation of this outcome would (most likely) refer to natural selection.

cognitive level[11] for this kind of situation. It does not rule out that a modified or refined experiment might indicate the presence of other problem solving mechanisms, e.g., one using different cues or cues of different intensity.

Notice that if the proposed setting would result in something like (A) and the same kind of animal, i.e., another individual of the same species, or, e.g., of the same cast, social status, or developmental stage, would also show judgment bias in the judgment bias test, that would still not mean that the animals do not misrepresent the ambiguous cue in the judgment bias test. It would, however, imply that for this kind of animal it is possible not to misrepresent the ambiguous cue and to represent it as conflicting. On the other hand, if the result would be something like (C) but the same kind of animals would show judgment bias in the judgment bias test, then this would strongly suggest that the animals in the judgment bias test are misrepresenting the ambiguous cue. The reverse, however, does not hold. If the animals are misrepresenting the ambiguous cue in the judgment bias test, it would not necessarily mean, in our setting, that they do not possess the relevant problem solving mechanisms.

4 Conclusion

Judgment bias tests allow assessing emotional states of non-human animals. Central to these tests is confronting animals with ambiguous cues that are intermediates between cues they have learned to link to positive and negative consequences, respectively, and to act accordingly. The mechanism of decision-making is usually taken to be a black box. We discussed how this black box could be opened, at least a little bit, even by experiments of the considered type. Drawing on the philosophical perspective of understanding decision-making as a capacity of certain representational systems, we determined three different ways that ambiguous stimuli could in principle be represented: *misrepresentation*, *conflicting content*, and *novel representation*. We judge misrepresentation to be the most likely scenario. Misrepresentation, however, does not imply the involvement of higher cognitive abilities that evaluate representations against each other. We propose a test regime in which the ambiguous stimulus is replaced by an ambiguous pair of unambiguous stimuli. This test regime makes it less likely that the animals misrepresent the ambiguous situation and aims primarily at testing the involvement of certain problem solving mechanisms that resolve a representation with a conflicting content. Finding out which species have this kind of mechanism would not only be

[11] There still might be mechanisms merely at the neuronal level to prevent "cognitive-freezing" and to cause the animals to get past the situation.

an interesting result in itself, but also help better understand the mechanism of biased judgment in non-human animals, which could help further develop judgment bias tests.

References

Adolphs R, Andler D (2018) Investigating emotions as functional states distinct from feelings. Emotion Review 10(3):191–201. https://doi.org/10.1177/1754073918765662

Agar N (1993) What do frogs really believe? Australas J Philos 71:1–12

Ahloy-Dallaire J, Espinosa J, Mason GJ (2018) Play and optimal welfare: Does play indicate the presence of positive affective states? Behav Processes 156:3–15

Barrett LF, Lindquist KA, Bliss-Moreau E, Duncan S, Gendron M, Mize J, Brennan L (2007) Of mice and men: Natural kinds of emotions in the mammalian brain? A response to Panksepp and Izard. Perspectives on Psychological Science: A Journal of the Association for Psychological Science 2(3):297–312. https://doi.org/10.1111/j.1745-6916.2007.00046.x

Boissy A, Manteuffel G, Jensen MB, Moe RO, Spruijt B, Keeling LJ, ... Aubert A (2007) Assessment of positive emotions in animals to improve their welfare. Physiology & Behavior 92(3):375–397. https://doi.org/10.1016/j.physbeh.2007.02.003

Bourin M (2015) Animal models for screening anxiolytic-like drugs: A perspective. Dialogues in Clinical Neuroscience 17(3):295–303. https://doi.org/10.31887/DCNS.2015.17.3/mbourin

Briefer EF, McElligott AG (2013) Rescued goats at a sanctuary display positive mood after former neglect. Applied Animal Behaviour Science 146:45–55. https://doi.org/10.1016/j.applanim.2013.03.007

Douglas C, Bateson M, Walsh C, Bédué A, Edwards SA (2012) Environmental enrichment induces optimistic cognitive biases in pigs. Appl Anim Behav Sci 139:65–73. https://doi.org/10.1016/j.applanim.2012.02.018

Dretske FI (1986) Misrepresentation. In: Bogdan RJ (ed) Belief: Form, content and function. Oxford University Press, New York, pp 17–36

Ekman P (1992) Are there basic emotions? Psychological Review 99(3):550–553. https://doi.org/10.1037/0033-295X.99.3.550

Godfrey-Smith P (1989) Misinformation. Can J Philos 19:533–550

Gygax L (2014) The A to Z of statistics for testing cognitive judgement bias. Anim Behav 95:59–69. https://doi.org/10.1016/j.anbehav.2014.06.013

Harding EJ, Paul ES, Mendl M (2004) Animal behavior – Cognitive bias and affective state. Nature 427:312

Hintze S, Melotti L, Colosio S, Bailoo JD, Boada-Saña M, Würbel H, Murphy E (2018) A cross-species judgement bias task: Integrating active trial initiation into a spatial go/no-go task. Scientific Reports 8(1):5104. https://doi.org/10.1038/s41598-018-23459-3

Izard CE (2010) The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. Emotion Review 2(4):363–370. https://doi.org/10.1177/1754073910374661

Jardim V, Verjat A, Féron C, Châline N, Rödel HG (2021) Is there a bias in spatial maze judgment bias tests? Individual differences in subjects' novelty response can affect test results. Behavioural Brain Research 407:113262. https://doi.org/10.1016/j.bbr.2021.113262

Kalueff AV, Stewart AM, Song C, Berridge KC, Graybiel AM, Fentress JC (2016) Neurobiology of rodent self-grooming and its value for translational neuroscience. Nature Reviews. Neuroscience 17(1):45–59. https://doi.org/10.1038/nrn.2015.8

Kremer L, Klein Holkenborg SEJ, Reimert I, Bolhuis JE, Webb LE (2020) The nuts and bolts of animal emotion. Neurosci Biobehav Rev 113:273–286. https://doi.org/10.1016/j.neubiorev.2020.01.028

Lagisz M, Zidar J, Nakagawa S, Neville V, Soroto E, Paul ES, Bateson M, Mendl M, Løvlie H (2020) Optimism, pessimism and judgement bias in animals: A systematic review and meta-analysis. Neurosci Biobehav Rev 118:3–17. https://doi.org/10.1016/j.neubiorev.2020.07.012

LeDoux J (2012) Rethinking the emotional brain. Neuron 73(4):653–676. https://doi.org/10.1016/j.neuron.2012.02.004

Mason GJ, Latham NR (2004) Can't stop, won't stop: Is stereotypy a reliable animal welfare indicator? Anim Welfare 13, Supplement 1:57–69.

Mathews A, Mogg K, Kentish J, Eysenck M, (1995) Effect of psychological treatment on cognitive bias in generalized anxiety disorder. Behaviour Research and Therapy 33: 293–303.

Mendl MT, Burman OHP, Parker RMA, Paul ES (2009) Cognitive bias as an indicator of animal emotion and welfare: Emerging evidence and underlying mechanisms. Appl Anim Behav Sci 118:161–181.

Mendl MT, Paul ES (2016) Bee happy: Bumblebees show decision-making that reflects emotion-like states. Science 353(6307): 1499-1500. https://doi.org/10.1126/science.aai9375

Millikan RG (2000) On clear and confused ideas: An essay about substance concepts. Cambridge University Press, New York.

Mineka S., Sutton SK. (1992) Cognitive biases and the emotional disorders. Psychological Science 3: 65–69.

Neville V, Nakagawa S, Zidar J, Paul ES, Lagisz M, Bateson M, Løvlie H, Mendl M (2020) Pharmacological manipulations of judgement bias: A systematic review and meta-analysis. Neurosci Biobehav Rev 108:269–286. https://doi.org/10.1016/j.neubiorev.2019.11.008

Nogueira SSdC, Fernandes IK, Costa TSO, Nogueira-Filho SLG, Mendl M (2015) Does trapping influence decision-making under ambiguity in White-Lipped Peccary (*Tayassu pecari*)? PLOS ONE 10. https://doi.org/10.1371/journal.pone.0127868

Paul ES, Harding EJ, Mendl M (2005) Measuring emotional processes in animals: The utility of a cognitive approach. Neurosci Biobehav Rev 29:469–491

Paul ES, Mendl MT (2018) Animal emotion: Descriptive and prescriptive definitions and their implications for a comparative perspective. Applied Animal Behaviour Science 205: 202–209. https://doi.org/10.1016/j.applanim.2018.01.008

Parker RMA (2008) Cognitive bias as an indicator of emotional state in animals. Unpublished PhD Thesis. University of Bristol.

Proctor HS, Carder G, Cornish AR (2013) Searching for animal sentience: A systematic review of the scientific literature. Animals 3(3):882–906. https://doi.org/10.3390/ani3030882

Ralph CR, Tilbrook AJ (2016) The usefulness of measuring glucocorticoids for assessing animal welfare. J Anim Sci 94:457–470

Richter SH, Kästner N, Kriwet M, Kaiser S, Sachser N (2016) Play matters: The surprising relationship between juvenile playfulness and anxiety in later life. Anim Behav 114:261–271

Richter SH, Schick A, Hoyer C, Lankisch K, Gass P, Vollmayr B (2012) A glass full of optimism: Enrichment effects on cognitive bias in a rat model of depression. Cogn Affect Behav Neurosci 12:527–542. https://doi.org/10.3758/s13415-012-0101-2

Roelofs S, Boleij H, Nordquist RE, van der Staay FJ (2016) Making decisions under ambiguity: Judgment bias tasks for assessing emotional state in animals. Front Behav Neurosci 10:119. https://www.doi/10.3389/fnbeh.2016.00119

Russell JA, Barrett LF (1999) Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. Journal of Personality and Social Psychology 76(5):805–819. https://doi.org/10.1037/0022-3514.76.5.805

Salmeto AL, Hymel KA, Carpenter EC, Brilot BO, Bateson M, Sufka KJ (2011) Cognitive bias in the chick anxiety–depression model. Brain Res 1373:124–130. https://doi.org/10.1016/j.brainres.2010.12.007

Scarantino A, Sousa R. de (2021) Emotion. In: Edward N. Zalta (ed) The Stanford encyclopedia of philosophy, (2021 ed). Metaphysics Research Lab, Stanford University

Simola N, Granon S (2019) Ultrasonic vocalizations as a tool in studying emotional states in rodent models of social behavior and brain disease. Neuropharmacology 159:107420. https://doi.org/10.1016/j.neuropharm.2018.11.008

Sterelny K (1990) The representational theory of mind: An introduction. Basil Blackwell, Cambridge, MA

Wang Q, Timberlake Matthew A, 2nd, Prall K, Dwivedi Y (2017) The recent progress in animal models of depression. Progress in Neuro-Psychopharmacology & Biological Psychiatry 77:99–109. https://doi.org/10.1016/j.pnpbp.2017.04.008

Whittaker AL, Barker TH (2020) A consideration of the role of biology and test design as confounding factors in judgement bias tests. Applied Animal Behaviour Science 232:105126. https://doi.org/10.1016/j.applanim.2020.105126