Lagisz, M., Zidar, J., Nakagawa, S., Neville, V. M., Sorato, E., Paul, E. S., Bateson, M., Mendl, M. T., & Løvlie, H. (2020). Optimism, pessimism and judgement bias in animals: a systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews*, *118*, 3-17. https://doi.org/10.1016/j.neubiorev.2020.07.012

Optimism, pessimism and judgement bias in animals: a systematic review and meta-analysis

Lagisz, Malgorzata [1†], Zidar, Josefina[2,†], Nakagawa, Shinichi[1,†*], Neville, Vikki[3], Sorato, Enrico[2], Paul, Elizabeth S.[3], Bateson, Melissa[4], Mendl, Michael[3#], Løvlie, Hanne[2#]

Addresses:

[1] Evolution and Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales, Sydney, NSW 2052, Australia

[2] The Department of Physics, Chemistry and Biology, IFM Biology, Linköping University, SE-581 83 Linköping, Sweden

[3] Centre for Behavioural Biology, Bristol Veterinary School, University of Bristol, Langford, BS40 5DU, United Kingdom

[4] Centre for Behaviour and Evolution, Biosciences Institute, Newcastle University, Newcastle upon Tyne, NE2 4HH, United Kingdom

*__Correspondence__: Shinichi Nakagawa s.nakagawa@unsw.edu.au and Michael Mendl Mike.Mendl@bristol.ac.uk

[†] These authors contributed equally to this work

[#] These authors supervised this work equally and are joint senior authors

Author contributions: two groups of the authors, SN & HL and MM, EP & MB conceived the idea independently, SN developed study design and methods from the inputs from others. JZ, ES, VN

and EP collected the data with inputs from SN, ML, MB, MM and HL. ML, JZ and SN conducted the analysis with inputs from MM. JZ, ML, SN, VN and MM co-wrote the first draft and all contributed to revisions of the manuscript.

Declarations of interest: none

Abstract

Just as happy people see the proverbial glass as half-full, 'optimistic' or 'pessimistic' responses to ambiguity might also reflect affective states in animals. Judgement bias tests, designed to measure these responses, are an increasingly popular way of assessing animal affect and there is now a substantial, but heterogeneous, literature on their use across different species, affect manipulations, and study designs. By conducting a systematic review and meta-analysis of 459 effect sizes from 71 studies of non-pharmacological affect manipulations on 22 non-human species, we show that animals in relatively better conditions, assumed to generate more positive affect, show more 'optimistic' judgements of ambiguity than those in relatively worse conditions. Overall effects are small when considering responses to all cues, but become more pronounced when non-ambiguous training cues are excluded from analyses or when focusing only on the most divergent responses between treatment groups. Task type (go/no-go; go/go active choice), training cue reinforcement (reward-punishment; reward-null; reward-reward) and sex of animals emerge as potential moderators of effect sizes in judgement bias tests.

Keywords: research synthesis, affective state, cognitive bias, animal welfare

Introduction

Accurate assessment of affect (emotion) in non-human animals is an important goal in disciplines including animal welfare science, neuroscience, psychopharmacology and drug development. A prevailing view in the study of human emotion is that affective states comprise subjective, behavioural, neural and physiological components (Paul et al., 2020; Scherer, 2005). Whilst the subjective component of animal affective states (feelings) is not currently accessible to direct measurement and we cannot be certain which species consciously experience such states (see Paul et al., 2020), we can objectively assess the other components. In his book *The Expression of Emotions in Man and Animals*, Darwin (1872) focused on behavioural manifestations of animal emotion, namely "expressive movements of the face and body"", and such measures continue to be used as indicators of animal affect today (e.g. Girard & Bellone, 2020). But other measures focus more directly on the role of affect in behavioural control and decision-making. A relatively new and promising approach is to measure biases in decision-making under ambiguity as indicators of animal affect (Harding et al., 2004; Mendl et al., 2009). This is because there are empirical and theoretical reasons to expect that responses to such ambiguity reflect affective valence (positivity or negativity of an affective state). For example, people in negative states are more likely to make negative ('pessimistic') judgements about ambiguous events or stimuli than people in more positive states (Blanchette and Richards, 2010; Paul et al., 2005). Such assessments could reflect an adaptive use of background affect (or mood) as a Bayesian prior over the likelihood of future positive or negative outcomes (Mendl et al., 2010; see Mendl & Paul, 2020 for a fuller discussion).

In line with these findings and ideas, a generic assay for measuring these so-called '*judgement biases*' has been developed for animals and has now been used in a large number of studies across a range of species. The original assay (Harding et al., 2004) involves training subjects to make one response (positive response) to a 'positive' cue (a single frequency tone) in order to

achieve a positive outcome (e.g. food) and a different response (negative response) to a 'negative' cue (a tone of a different frequency) in order to avoid a negative outcome (e.g. white noise) (Figure 1a). Once subjects have learnt this conditional discrimination task, training continues but includes occasional ambiguous cues (tones of intermediate frequency) designed to assess whether subjects would make the positive response indicating anticipation of a positive outcome, or the negative response indicating anticipation of a negative outcome. This allows one to test whether, for example, animals in a putative negative affective state (e.g. as a result of some sort of experimental treatment, Figure 1b) are more likely to make the negative response, as predicted (Figure 1c,d). Making the positive or negative response under ambiguity can be operationally defined as 'optimistic' or 'pessimistic' (Bateson, 2016) without implying that animals experience optimism or pessimism as humans do.

Published studies using this judgement bias task (also referred to as an 'ambiguous cue interpretation' task (Rygula et al., 2013)), and variants of it, have supported the general prediction, but also generated null and opposite results. These findings have been summarised narratively in a number of review papers that have also identified various methodological and theoretical questions regarding the task and approach (Baciadonna and McElligott, 2015; Bethell, 2015; Gygax, 2014; Hales et al., 2014; Mendl et al., 2009; Mendl & Paul 2020; Roelofs et al., 2016). What has been lacking, and much needed, is a systematic review and meta-analysis of the findings to date to evaluate whether the general predictions behind the approach are supported, and how results may be influenced by a variety of moderators, including aspects of task design, methods used to manipulate affect, species studied, and age and sex of subjects. Recently, we published the first such meta-analysis focusing on the effects of pharmacological manipulations of affective state on judgement biases (Neville et al., 2020). Here we systematically review and meta-analyse the much larger number of studies that have used non-pharmacological affect manipulations.

We focus on judgement bias tasks based on the Harding et al. (2004) method, since these have been more widely studied in animals than other cognitive biases such as attention (Bethell et al., 2016; Crump et al., 2018) and memory biases (Burman and Mendl, 2018). Although details of the procedures and criteria used to select suitable studies and extract appropriate data for the meta-analysis are explained in the Methods section, three points should be noted here.

First, a major challenge in any study of animal affect is to establish a 'ground truth' for the affective state that the animal is in when under study. This is necessary, for example, if an aim of a study is to determine what behavioural, physiological or neural changes occur in animals in particular affective states, and hence to develop reliable indicators of such states. Therefore, in studies which seek to evaluate whether judgement bias is a valid indicator of affective valence, we need to know whether the animal is in a relatively positive or negative state, so that we can test whether animals that are in a more positive state do indeed show more optimistic decisions under ambiguity, than those in a more negative state. In most judgement bias studies, researchers attempt to use an experimental treatment to induce a relatively positive or negative affective state compared to a control or 'benign' treatment group, or they impose both a positive and a negative treatment, and compare these. Because we cannot know for certain where the intermediate 'neutral' state lies, we use terminology that emphasises the relative nature of these manipulations. Thus, we refer to 'better' (more positive), 'benign/control', and 'worse' (more negative) treatments, and assign them to either 'relatively better' or 'relatively worse' groups for pair-wise comparison in the meta-analysis.

Second, there are two main types of task used in judgement bias trials: active choice (go/go) and go/no-go (Figure 1a). In go/go active choice tasks, the animal has to choose between two alternative responses (e.g. press the left or right lever), while in go/no-go tasks the animal's options are to perform a response (e.g. approach a location or press a lever) or supress it. The response of animals can be reported as a proportion (e.g. proportion of trials in which the

subject pressed the left lever or approached the location), or a latency (e.g. time taken to press the lever or approach the location). Latency and proportion data have different statistical distributions; they require different transformations and use of different formulae to calculate effect sizes. There are also biological reasons for separating latency and proportion data. Different measures may represent different aspects of cognitive processes and their utility depends on the type of cognitive bias task used. In go/no-go tasks, latency to perform the response under ambiguity is a direct measure of judgement bias. For example, if the positive response is to approach the cue, then quick approach to an ambiguous cue indicates an 'optimistic' response. In contrast, go/go active choice tasks require responses to both cues (e.g. press left or right lever), meaning that the latency to perform whichever response the animal selects is more difficult to interpret in terms of 'optimism' or 'pessimism'. Rather, the proportion of positive or negative responses provides more definitive information about 'optimistic' or 'pessimistic' decisions. This measure is also of use in go/no-go tasks. Therefore, proportion of positive vs. negative responses is preferable to latency as a measure of judgment bias for go/go active choice tasks, whereas for the go/no-go tasks both measures are, in principle, suitable.

Third, many judgement bias studies use more than one ambiguous cue during test trials (Figure 1c). Often three such cues are used; one (MID) which is assumed to be perceived by the animal as being at the mid-point of the sensory scale (e.g. sound frequency) between the positive (P) and negative (N) training cues, one (near positive: NP) which is half way between MID and the positive (P) cue, and one (near negative: NN) which is halfway between MID and the negative (N) cue. There are theoretical and methodological reasons for why an affect manipulation treatment might have an influence at one ambiguous cue but not at others in the same study. For example, non-midpoint ambiguous cues (NP, NN) may be perceptually too similar to the P and N training cues for animals to moderate their responses to them, whilst the midpoint (MID) cue is usually ambiguous enough for background affect to influence responses to it. In some studies

MID could be perceived as closer to P or N and the most ambiguous cue becomes either NN or NP, respectively. Moreover, the perceived payoff of the positive and negative response outcomes, and hence associated decisions, may be asymmetrical. For example, if the perceived negative value of a foot-shock outcome is much stronger than the perceived positive value of a food pellet, animals may be strongly motivated to avoid shock risk and thus respond negatively to both MID and NN ambiguous cues, with variation in response limited to the 'safest' NP ambiguous cue (Mendl et al., 2009). Conversely, in a test variant where negative cues are simply lacking a reward instead of bearing a punishment, animals may respond positively even to negative cues, because the cost of doing so is negligible. Because it is likely that biased responses are unevenly spread across ambiguous cues – in fact some studies report effects which are strongest or only statistically significant at one ambiguous cue location (e.g. Bethell and Koyama, 2015; Zidar et al., 2018) – we investigate the effect of relative cue position and also conduct sensitivity analyses. These additional analyses use data subsets with different decision rules for selecting the most representative data points from response curves (e.g. using only the ambiguous cue with the largest absolute between-treatment effect size; more details in the Methods section).

This systematic review and meta-analysis aims to: (i) quantify the overall effect size that affect manipulations have on measures of judgement bias in animals; (ii) estimate heterogeneity of the results among different studies; (iii) explore the influences of different biological and methodological moderators (explanatory variables for variation in effect sizes).

Methods

Literature search

We conducted a systematic literature search and recorded relevant information required in the

Preferred Reporting Items for Systematic reviews and Meta-Analyses statement (PRISMA; Moher et al., 2009; see Supplementary Materials for additional search details). We ran the first online database search on 29 October 2015, a second search in December 2017, and a final database search to update the dataset again on 27 March 2019. For these searches, we used the broad-coverage interdisciplinary databases *Scopus* and *Web of Science*, covering the titles, abstracts and keywords of academic publications.

The initial search string used in *Scopus* was: TITLE-ABS-KEY (("cognitive bias*" OR "judgment bias*" OR "judgement bias*" OR "cognitive affective bias*") AND (pessimis* OR optimis* OR valence OR mood* OR emotion* OR "affective state*" OR "emotional state*" ambig* OR animal* OR "animal welfare")) AND PUBYEAR > 2003 and in *Web of Science*: TS=(("cognitive bias*" OR "judgment bias*" OR "judgement bias*" OR "cognitive affective bias*") AND (pessimis* OR optimis* OR valence OR mood* OR emotion* OR "affective state*" OR "emotional state*" ambig* OR animal* OR "animal welfare")) AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article), Indexes=SCI-EXPANDED, SSCI Timespan=2004-2015. We restricted the publication years to those following the seminal paper on animal judgement bias (Harding et al., 2004). We restricted the subsequent updates of the literature search to the years since the previous search update (i.e. 2015-2017 and 2017-2019, respectively) and otherwise used the same search strings. We collected additional relevant studies from the authors whom we contacted to request data or other additional information that was missing from their publications. We also performed searches of reference lists of relevant review articles and research articles citing the seminal study by Harding et al. (2004).

The searches of the online databases generated over 900 potential article references and searches of other sources generated almost 500 additional references for screening (Figure 2). We removed duplicated results from these separate search paths. Two authors (J.Z. and E.S.) independently screened 482 abstracts from the articles identified in the 2015 search using the

software AbstrackR (Wallace et al., 2012). M.L. performed two updates of literature searches in 2017 and 2019, following the same methodology as in the first search. Overall, we identified 74 published studies as potentially suitable for inclusion in our meta-analysis after screening of full texts and removal of duplicated studies. We excluded three studies during the data extraction stage (due to missing data), resulting in data from 71 studies being included in the meta-analysis.

Inclusion and exclusion criteria

We screened titles and abstracts from bibliometric records to identify empirical studies on judgement bias in animals in which subjects were exposed to an affect manipulation aimed at inducing either a relatively positive or negative state. We then screened full text versions of the articles that passed this initial screening stage. At the full-text screening stage, the following six criteria had to be met for the study to be included in the meta-analysis: i) study had to be experimental and designed to investigate variation in judgement bias (i.e. 'optimistic' or 'pessimistic' interpretation of stimuli) in non-human animals; ii) experiments had to include at least two treatment groups (or control/'benign' and treatment groups); iii) experimental treatments had to be designed to induce 'relatively better' or 'relatively worse' affective states (see decision-tree in Supplementary Materials Figure S1); iv) for go/no-go tasks studies had to report either latency to make a response to ambiguous cues, or proportion of go or no-go responses towards ambiguous cues; for active choice tasks, studies had to report proportion of positive or negative responses; if the data available could be translated into such latencies or proportions, they were included; v) studies had to present data usable for effect size calculation; if suitable data could not be retrieved by contacting the authors, the study was excluded from the meta-analysis; vi) studies had to be published in peer-reviewed journals, but student reports and data from unpublished work, as well as articles that were written in languages other than English, could have been included if they met the above criteria.

We also excluded studies for the following additional reasons. We only considered data from studies investigating judgement bias, i.e. we excluded studies investigating other cognitive biases, such as attention bias and memory bias. We also excluded studies only describing judgement bias theory or methods or reviewing previous findings and studies that used the generic judgement bias task for humans, because our focus was on non-human animals. As studies investigating effects of drugs on judgement bias often include several doses that cannot easily be assigned into relatively better and relatively worse treatment groups, we also excluded all drug studies from this meta-analysis. As mentioned earlier, the drug studies were recently subjected to a separate meta-analysis by our group (Neville et al., 2020).

Data extraction

After compiling a final list of included studies, we extracted measurements representing behavioural responses to cues in the judgement bias tests. Each pairwise comparison consisted of a pair of outcome measures comparing behaviour of animals from 'relatively better' to 'relatively worse' affect manipulation groups. Our classification of treatments as inducing 'relatively better' or 'relatively worse' affective states was based on a decision tree involving screening articles and assessing treatments based on the following three criteria. First, if stated, we used the *a priori* hypothesis and reasoning outlined in the research article. Second, where possible, we employed Rolls' (2005, p.11) operational definition of emotion as "states elicited by rewards and punishers", where "a reward is anything for which an animal will work" and "a punisher is anything that an animal will work to escape or avoid". Thus, if a treatment involved stimuli that the subject animal is known to actively avoid, we deemed it to induce a relatively worse affective state than one which involved neutral or preferred stimuli. Third, we considered evidence from previous studies on the effects of the treatments in question on affective state (e.g. their effects on other putative indicators of affective state, such as abnormal repetitive behaviour or physiological stress indicators).

The decision tree for assigning affect treatments to relative affect manipulation categories is presented in Supplementary Materials Figure S1. If the first criterion in the decision tree was fulfilled (i.e. the authors of the original paper explicitly stated whether the treatment is expected to have positive/negative effect on animals' affective state), we ignored the subsequent decision criteria. If not, we evaluated the subsequent decision criteria. We classified all extracted treatment groups within a study relative to each other. For example, in a study with a control (benign/unmanipulated) and enriched housing group, the enriched group would be considered 'relatively better' and the control/benign group 'relatively worse'. Conversely, in a study with control/benign and stress-induction groups, the stress group would be considered 'relatively worse', and the control/benign group would be considered 'relatively better'.

We tackled variation in study design and outcome measurement as follows. First, for the go/no-go judgement bias tasks we extracted either or both (depending on which was reported) latency and proportion outcome measures (the signs of the effect sizes calculated from latency measurements were later inverted, so that interpretation of the effect direction was consistent with that for the proportion data). For active choice go/go judgement bias tasks, we extracted only proportion outcome measures (as explained earlier, latency measures in active choice tasks cannot be clearly linked to more 'optimistic' or 'pessimistic' responding). The extracted mean and standard error (or standard deviation) of responses to ambiguous and non-ambiguous cues during the tests were used to calculate values of effect sizes (and their variances) for each pairwise comparison of the relatively better and relatively worse treatment groups at the same cue. Relevant sample sizes were also recorded representing the number of animals from each group participating in the judgement bias test.

Second, included studies used varying numbers of ambiguous cues (range 1-13, mean 2.99, mode 3). We only extracted data for a maximum of three ambiguous cues per measurement (response curve). We always extracted data for the middle cue (midpoint between the positive

and negative cues, MID) and, if available, two intermediate cues between the middle cue and positive and negative cues (near-positive NP and near-negative NN, respectively). If response data to positive (P) and negative (N) cues were reported for judgement bias tests, these were also extracted.

Third, when judgement bias was measured on several consecutive days following a treatment, we extracted the first measure only as it was usually closest in time to the acute affect manipulation treatment (Destrez et al., 2013; Doyle et al., 2011). In a few studies, animals were exposed to several judgement bias tests during a long-term treatment (Douglas et al., 2012; Hales et al., 2016; Rygula et al., 2013). In these cases, we extracted the last test occurring during each treatment, thus maximising the time available for it to exert its effects. We assumed that the cumulative impact of chronic exposure to the treatment likely out-weighed any potential effect of learning about repeated tests that were spaced out across time.

Fourth, some studies with a within-subject design measured judgement bias before, during and after an affect manipulation (e.g. pre-stress, stress, post-stress), or repeated the 'baseline' treatment (e.g. enriched, barren, enriched) (Brilot et al., 2010; da Cunha Nogueira et al., 2015; Hales et al., 2016; Murphy et al., 2013). In these studies, we compared measures taken before treatment ('baseline') to those taken during it and did not include measures taken after it.

Fifth, studies using a between-subject design sometimes tested both control/benign and treatment groups before, during and after a manipulation. In these cases, we compared the control/benign group to the treatment group during treatment and ignored the pre- and post-treatment measurements (Hales et al., 2016; Oliveira et al., 2016; Rygula et al., 2013).

Finally, if several treatments were applied where one or more treatments were hypothesized to be intermediate in effect to the two most extreme treatments, only the two extreme treatments

were included (Ash and Buchanan-Smith, 2016; Burman et al., 2009; Keen et al., 2014; Wheeler et al., 2015).

For each experiment, we gathered information on the potential moderator variables to characterise our dataset and explain potential heterogeneity in the data. Detailed descriptions of all the originally extracted moderators are included in Table S1. In brief, the three key groups of extracted moderators considered information about the article, biological variables, and test design. Paper-specific information included authors, title, journal, and publication year. For each data point (i.e. comparison between two groups of animals), we extracted the following biological variables: taxa studied (mammals, birds, insects), sex (female, male, mixed-sex), age class (juvenile, adult) and source of animals (captive, wild-caught). Test-specific information included affect manipulation category (enrichment, stress, other), affect manipulation timing (before/during test, long-term), comparison category (Better-Worse, Benign-Worse, Better-Benign), type of cue used in judgement bias test (spatial, visual, auditory, tactile, olfactory), whether animals were food deprived prior to behavioural trials (yes, no/no information), automation of response measurement (yes, no/no information), blinding of personnel performing trials (yes, no/no information), combination of reinforcement used during training (Reward Vs. Null, Reward Vs. Punishment; Reward Vs. Smaller Reward), task type (active choice go/go, go/no-go), whether ambiguous cues were reinforced (yes, no/no information), measurement type (latency, proportion), location of ambiguous cues relative to positive and negative cues (P – positive, NP – near-positive, MID – midpoint, NN – near-negative, N – negative). We also noted any pertinent additional details about study designs (between-subjects, within-subjects), affect manipulations, source of the data in the original studies, and any associated comments. When data were provided in a graph instead of a table or text, we extracted the values using GraphClick 3.0.3 (http://www.arizona-software.ch/graphclick/). Data extraction was performed by J.Z., M.L. and V.N. and was checked by M.L., V.N. and E.S.

Effect-size calculation

We used Hedges' unbiased standardized mean difference (Hedges' g) as the measure of effect size. Because latency and proportion data are bounded (i.e. latencies start at 0 and are often censored, and percentages are bounded between 0 and 100, proportions between 0 and 1), we used natural log (for latencies) or logit-transformed data (for proportions, and percentages expressed as proportions) to calculate Hedges' g (details provided in Supplementary Materials Methods and Figure S2). In brief, to calculate Hedges' g, we focused on positive responses (i.e. those which indicated that the subject was anticipating a more rewarding outcome) and subtracted the mean value of the relatively worse treatment from the mean of the relatively better treatment, and divided the difference by the pooled standard deviation (SD) with correction for small sample sizes (Hedges and Olkin, 1985). Thus, if animals from the relatively better treatment group were making a higher proportion of positive responses than animals from the relatively worse treatment, the difference between the means would be positive and the effect size too. However, the expected pattern would be reversed when latencies to make the positive response were measured in go/no-go tasks: if animals from the relatively better treatment group were quicker to make the positive response (i.e. had lower latencies), than animals from the relatively worse treatment, the difference between the means (and the resulting effect size) would be negative. To allow for easier comparison and interpretation of the effect sizes from latency and proportion measures, we reversed the sign of the effect sizes based on latency measures. Thus, after the sign adjustment, across all data positive values of Hedges' g can be interpreted as optimistic responses of animals exposed to relatively better treatments compared to those exposed to relatively worse treatments. For the go/no-go tests that reported the outcomes as both latency and proportion, we calculated Pearson's correlation between these two measures.

Meta-analysis and meta-regression models

We ran all statistical analyses in *R* version 3.6.0 (R Development Core Team, 2019); we created main forest-like (orchard) plots of effects using *orchaRd* package (Nakagawa et al., 2020). For multilevel meta-analysis and meta-regression we used the *rma.mv* function from the package *metafor* (Viechtbauer, 2010).

To estimate the overall mean of the effect sizes we constructed intercept-only models (i.e. meta-analysis) with study ID, experiment ID, cue ID, and effect size ID as random effects. To explore effect of species identity and phylogenetic relatedness, we also evaluated meta-analytic models with phylogeny and species ID added to the random effects list. We calculated I^2 values for each random factor and the overall heterogeneity, I^2_{Total}, in the meta-analytic models (Nakagawa and Santos, 2012).

To evaluate the effects of moderators of interest (e.g. subject sex or age class, test task type, test cue type and level of cue ambiguity), we ran univariate multilevel phylogenetic meta-regression models with moderators as fixed effects, and the same random effects as in the meta-analytic models (except species ID). In the multivariate meta-regression models (i.e. models with multiple moderators), we included only moderators that were significant in the univariate meta-regression models. We then performed AICc-based model selection using MuMIn package (Barton, 2009) to infer relative contributions of included moderators. To assess the fit of meta-regression models, we calculated marginal R^2 values (*sensu* Nakagawa and Schielzeth, 2013; Nakagawa et al., 2017).

Publication bias

Statistically significant results are more likely to be published, resulting in a non-random sample of data available for meta-analysis (Rosenthal, 1979). To examine publication bias in our data set, we visually inspected a funnel plot for asymmetry in the distribution of the residuals of effect sizes (which are the sum of effect size level effects and sampling variance effects; i.e. meta-

analytic residuals: *sensu* Nakagawa and Santos, 2012). We also performed Egger's regression on the residuals and measurement errors from the full meta-regression model (multilevel version of the publication bias test; Nakagawa and Santos, 2012). Egger's regression indicates publication bias if the regression intercept is significantly different from 0 (Egger et al., 1997). Finally, we tested for a special type of publication bias, a time-lag bias, i.e. a tendency for studies with larger effects to be published earlier (Jennions and Møller, 2002).

Sensitivity analyses (robustness of results)

To test robustness of our results to the estimation method, we ran a meta-regression model and a multilevel mixed-effect full meta-regression model (with subject sex, task type, cue type, and reinforcement type as moderators), using a Bayesian approach, as implemented in the *MCMCglmm* package (Hadfield, 2010). These models were run with 110,000 iterations, 10,000 burn-in periods, and thinning by every 100 resulting in an effective sample size of 1000. We used a parameter-expanded prior ($V = 1$, $nu = 1$, $alpha.mu = 0$, $alpha.V = 1000$), with EffectID (units) fixed at one.

We also ran the meta-analytic models using four additional data configurations representing different ways of interpreting results from pairs of response curves with multiple cues tested. First, we used a dataset with positive and negative test cues excluded, so that only responses to ambiguous cues were used (maximum of 3 effect sizes per comparison of pair of response curves: for near-positive, midpoint, near-negative cues). In the remaining data subsets, we selected only one cue per response curve comparison. Thus, to create the second data subset, we only included data from the mid-point ambiguous cue location (MID data points and effect sizes). In the third data subset, we selected the effect sizes data from the cue location with the largest absolute value within each response curve comparison; notably, in 71.3% of the comparisons, the largest absolute effect size was not located at the mid-point ambiguous cue. In

the fourth data subset, we used effect sizes with the biggest absolute value in the direction of the mean value, within each response curve comparison, as in Neville et al. (2020).

Results

Description of data set

The workflow and outcomes of our systematic literature searches are presented in a PRISMA diagram (Figure 2). The list of included studies is provided in Supplementary Table S2. Excluded studies, with reasons for exclusion, are listed in Supplementary Table S3. To retrieve missing data, or additional information, we contacted 39 authors about 35 studies. We attained raw data for 18 studies and additional information for 10 studies. Ultimately, we extracted 459 effect sizes, representing 91 experiments published in 71 articles. These studies were performed on 22 species, ranging from bees to monkeys. The main characteristics of the included studies are summarised in Figure 3, showcasing significant variation in study subjects and methodologies. Individual studies contributed between 1 and 30 effect sizes to our final data set.

Mammals were the best-represented taxonomic group (56 out of 71 studies; 330 out of 459 effect sizes), and almost all studies were performed on captive animals (65 studies; 414 effect sizes). Females were more frequently used in experiments than males or mixed-sex groups (225, 118, 116 effect sizes, respectively; for the numbers of studies see Figure 3), and adults were more commonly used than juveniles (333 and 126 effect sizes, respectively). Most often, affect manipulation was a form of stress induction compared to standard/benign conditions (benign-worse comparison: 230 effect sizes). Enrichment compared to control/benign conditions was the next most common manipulation (better-benign comparison: 135 effect sizes), and a few studies compared positive treatments (e.g. enrichment) to negative treatments (e.g. handling) (better-worse comparison: 94 effect sizes). Manipulations were usually long-term (292 effect sizes), lasting for days or weeks before affect was measured.

Between-subject designs (independent groups of animals exposed to manipulation or control/benign treatment) accounted for 302 effect sizes and within-subject designs accounted

for 157 effect sizes. Go/no-go tasks dominated over active choice go/go tasks (389 and 70 effect sizes, respectively). Spatial and visual cues were most commonly used in judgement bias tests (177 and 167 effect sizes respectively), and reward-punishment training schemes were more common than reward-null (283 and 132 effect sizes, respectively), with the remaining studies using different reward strengths (44 effect sizes). Most studies did not report whether the personnel performing measurements of animal behaviour were blinded to treatments (only 113 effect sizes came from blinded trials), or whether the measurements were automated (only 71 effect sizes came from automated trials). Finally, latency and proportion outcome measures were reported at similar levels (258 and 201 effect sizes, respectively). Only 5 studies using go/no-go tasks reported outcome measures as both latency and proportion, and these were moderately correlated (r = 0.578, t = 3.085, df = 19, p-value = 0.006), although not for the data subset using only the largest effect sizes from each experiment to remove non-independence (r = 0.443, t = 0.857, df = 3, p-value = 0.455).

An overall effect and heterogeneity among effect sizes

Overall, we found a statistically significant effect of experimental treatments on judgement bias in animals (phylogenetic multilevel meta-analysis: Hedges' g $(H_g)_{[\text{overall mean}]}$ = 0.201, 95% Confidence Interval (CI) = 0.028 to 0.374; Figure 4, Table S4). A similar model, but without controlling for phylogeny, also showed a statistically significant overall effect (multilevel meta-analysis: $H_{g[\text{overall mean}]}$ = 0.204, 95% CI = 0.087 to 0.320, Table S5). Therefore, animals in a relatively better treatment usually behaved in a more 'optimistic' way than animals in a relatively worse treatment, whereas animals in a relatively worse treatment were more 'pessimistic'. Notably, this overall effect is comparable to a small effect, as suggested by the benchmark values (0.2, 0.5 and 0.8 as small, medium and large effects; Cohen, 1969). The total heterogeneity in the whole data set was high (I^2_{total} = 76.4%; according to Higgins' benchmark 25, 50 and 75% can be interpreted as low, moderate and high heterogeneity, respectively;

Higgins and Thomson, 2002). About 68.1% of the variability across studies was due to sampling error, while phylogeny contributed little to account for this heterogeneity (2.0%), suggesting a weak phylogenetic signal (see Nakagawa & Santos, 2012).

High observed total heterogeneity in the data set warrants investigation of potential moderators of heterogeneity. We, thus, present findings of the univariate multilevel phylogenetic meta-regression models examining the effects of different moderators (see Figures 5, Figure 6, Figure S3).

Species-specific effects

A meta-regression model estimating mean effect for each included species did not show a clear pattern of differences among species (Figure 5; $R^2 = 0.070$, Table S6). Some of the species-specific point estimates were medium or large, but they were accompanied by wide confidence intervals crossing zero (no-effect) line. We note that the distribution of studies among species was not balanced, with the data set being dominated by studies on rats, cattle, and pig (15, 11 and 8 studies, respectively), while most of remaining species are each represented by a single study (Figure 5).

Sex-effects

Effects of judgement bias manipulations on males were small-to-medium and statistically different from zero ($H_{g[\text{males}]}$ = 0.365, 95% CI = 0.155 to 0.575), while effects on females were, on average, close to zero ($H_{g[\text{females}]}$ 0.104, 95% CI = -0.063 to 0.271; Figure 6a). The difference between mean effects in males and females was small ($H_{g[\text{male vs. female difference}]}$ = 0.261, 95% CI = -0.001 to 0.522; $R^2 = 0.024$, Table S7), indicating that affect manipulations on judgement bias measurements tend to be more pronounced in studies on males than females.

Tasks type effects

Effects of judgement bias manipulations tended to be larger in studies using active choice tasks in comparison to studies using go/no-go tasks ($H_{g[\text{go/no-go vs. active choice difference}]}$ = -0.277, 95% CI = -0.567 to 0.012; Figure 6b). On average, tasks with active choice had a medium effect size and were statistically different from zero ($H_{g[\text{active choice}]}$ = 0.432, 95% CI = 0.151 to 0.712), while the average effect size in in go/no-go tasks was small, but still statistically different from zero ($H_{g[\text{go/no-go}]}$ = 0.154, 95% CI = 0.005 to 0.304; $R^2 = 0.021$, Table S8).

Cue types used during judgement bias tests

Across the five categories of cues used during judgement bias tests, only tests using auditory and tactile cues consistently revealed differences between control and affect-manipulated groups of animals ($H_{g[\text{auditory cues}]}$ = 0.393, 95% CI = 0.136 to 0.651; $H_{g[\text{tactile cues}]}$ = 0.658, 95% CI = 0.136 to 1.118; Figure 6c). These two categories of cues were only significantly different from the results from studies using visual cues, which on averaged had the weakest effect ($H_{g[\text{visual cues}]}$ = 0.067, 95% CI = -0.133 to 0.268; $R^2 = 0.044$, Table S9).

Reinforcement scheme during judgement bias tests

Studies using Reward-Punishment and Reward-Reward training cue reinforcement schemes usually generated small-medium statistically significant effect sizes in the predicted direction ($H_{g[\text{Reward-Punishment}]}$ = 0.216, 95% CI = 0.036 to 0.396; $H_{g[\text{Reward-Reward}]}$ = 0.488, 95% CI = 0.137 to 0.839), but not the Reward-Null reinforcement scheme (Figure 6d). Reward-Reward studies generally showed significantly larger judgement bias than those that used a Reward-Null reinforcement scheme ($H_{g[\text{Reward-Reward vs. Reward-Null}]}$ = 0.436, 95% CI = 0.045 to 0.827; R^2 = 0.030, Table S10). Studies using non-reinforced ambiguous cues (which was the vast majority of included studies) generated effect sizes in the predicted direction ($H_{g[\text{ambig. cue not reinforced}]}$ = 0.204, 95% CI = 0.026 to 0.382), although not statistically different from studies in which ambiguous

cues were reinforced ($H_{g[\text{ambig. cue not reinforced vs. reinforced}]}$ = 0.080, 95% CI = -0.527 to 0.686; R^2 = 0.001), whose effect sizes were close to zero (Table S11).

Cue ambiguity level

Ambiguous cues that were halfway between the positive and negative cues, as well as cues that were closer to the negative cues, were most likely to reveal judgement bias in tested animals ($H_{g[\text{mid-point cue}]}$ = 0.250, 95% CI = 0.042 to 0.458; $H_{g[\text{near-negative cue}]}$ = 0.303, 95% CI = 0.075 to 0.530; R^2 = 0.014, Figure 6e). Ambiguous near-negative cues were also significantly different from the effects of positive training cues, with the latter on average being least likely to show judgement bias effect ($H_{g[\text{positive cues}]}$ = 0.063, 95% CI = -0.153 to 0.278, Table S12).

Other moderators in univariate models

Variation in the other considered moderators did not appear to significantly influence the magnitude of judgement bias effects. These moderators were: source of animals (captive vs. wild-caught), animal age, type of affect manipulation (stress vs. enrichment), timing of affect manipulation (short vs. long-term), whether manipulation was compared to benign or worse reference condition, type of study design (within-individual vs. between-individuals), food deprivation during judgement bias tests, measurement type of behavioural response (latency vs. proportion), automation and blinding of measurements of animal responses (Figure S3; Tables S13 – S22; R^2 = 0 to 0.010).

Multivariate (full) meta-regression models and model selection

The full meta-regression model included four moderators that were significant or close to statistical significance in univariate models (after confirming they were not co-linear with each other): sex of test animals, task type (go/no-go vs. active choice go/go), type of cue used in the test, and type of reinforcement for positive and negative training cues. In the multivariate meta-regression, none of the considered moderators was significant (Table S23). These moderators

can jointly explain only about 7% of variation in the data (R^2 = 0.072). Model selection analysis indicated that type of the task and type of reinforcement used could be the most influential moderators, followed by the sex of animals (Table S24).

Publication bias

We conducted 3 kinds of publication bias analyses: 1) contour-enhanced funnel plots of residuals, 2) a variant of Egger's regression, and 3) a regression-based time-lag bias test. Visual inspection of enhanced-contour funnel plots of residuals did not reveal skewness indicative of publication bias (Figure S4). Further, the intercept of Egger's multivariate regression, controlling for potentially important moderators from univariate models, was not significantly different from zero (t = 0.017, df = 457, p = 0.986), confirming lack of publication bias in the full data set. Finally, we found no evidence for time-lag bias, as the slope of linear regression between publication year and effect size was not significantly different from zero (Slope $_{[Year]}$ = -0.002, 95% CI = -0.121 to 0.118, p = 0.980, Table S25).

Sensitivity analyses (robustness of results)

The estimates from Bayesian models run on full data set gave qualitatively identical results to the REML models used in the main data analyses. Namely, the overall effect was small and statistically significant ($H_{g[overall\ mean]}$ = 0.206, 95% CI = 0.041 to 0.383; I^2_{total} = 76.8%; Table S26). In the Bayesian multivariate meta-regression, none of the moderators significantly influenced judgement bias test outcomes, as in the equivalent log-likelihood model.

Finally, we ran meta-analytic models on four data subsets, representing different ways of looking at the results from response curves with multiple cues: i) including only data from ambiguous cues (81 NP, 108 MID, and 80 NN effect sizes for cue locations included in this data subset), ii) including only data from mid-point ambiguous cues (108 MID effect sizes included), iii) including only data for maximum response, in absolute terms (26 P, 13 NP, 31 MID, 22 NN,

and 16 N effect sizes included), iv) including only data for maximum response in the overall direction of response (19 P, 12 NP, 34 MID, 28 NN, and 15 N effect sizes included). All these data subsets tended to have larger overall effect size estimates, than in the full data set meta-analyses (Figure 4, Tables S4 and S5). Univariate and multivariate meta-regression models usually showed similar patterns to these observed in the analyses on the full dataset (Tables S6-S23).

Discussion

Our meta-analysis revealed that non-pharmacological affect manipulations generally influenced judgement bias in the predicted direction (i.e. manipulations assumed to generate a relatively positive state were likely to generate an 'optimistic' response to cues). However, effects were usually small to large (average Hedges' g of 0.2 – 0.6), and they were highly variable, with total observed heterogeneity (I^2) over 75%. The moderators that potentially influenced magnitude of effects included cue type, type of task used in judgement bias trials, reinforcement combination used for training positive and negative cues, cue ambiguity level, and sex of tested animals. However, small R^2 values (1.4 to 4.4%) indicated that these moderators explained only a small proportion of variance. We discuss these findings in detail below.

Validity and efficacy of judgement bias tests

Our main finding generally supports judgement bias tests as a valid approach to measure affect in non-human animals. This is in line with conclusions of a narrative cross-species review (Bethell, 2015) and a recent systematic review of 20 rodent studies on judgement bias (Nguyen, et al. 2020). However, the latter considered both pharmacological and non-pharmacological manipulations and only conducted a qualitative synthesis of their rodent data set. Effects of pharmacological manipulations across species were recently quantitatively synthesised by our team (Neville, et al. 2020) and our current work provides the first quantification of non-pharmacological manipulations across different taxa.

Our quantitative results show that the observed behavioural effect of the affect manipulations investigated is, on average, small (Hedges' g of 0.2) and highly heterogeneous. However, we base this conclusion on the analyses of the full dataset, which included mean latency and/or proportion data from all cues used in the judgement bias tests. Thus, we likely underestimated the overall effect size, due to the inclusion of positive and negative training (unambiguous) cues,

in the analysis. Weak discrimination performance for the training cues could decrease the likelihood of detecting a judgement bias (Roelofs et al. 2016), warranting further investigations of how selected learning criteria, or actual discrimination levels, affect the sensitivity of judgement bias tests. It is also possible that the affect manipulations used in many of the included studies were rather "mild" – not many authors used severe stressors or pain stimuli, either for welfare reasons and/or because pain stimuli might exert a general suppressive effect on responses to cues. It is thus possible that some manipulations failed to influence animal affect.

As noted earlier, there are theoretical and empirical reasons for why judgement biases may not occur at training cues, and also for why they may not occur at all ambiguous cues. When we restricted analysis to the cue with the largest absolute effect size in the direction of the overall mean effect size from each response curve – the estimated overall effect sizes were between moderate to large (Hedges' g of 0.6). The overall effect sizes were moderate when we used the other three data subsets: (i) ambiguous cues only; (ii) middle cue only; (iii) cue with the largest absolute effect size. Yet, analyses of the full dataset are most powerful, given that they include data points representing the whole response curve (Gygax, 2014).

The high data heterogeneity is congruent with the levels observed in most ecological and evolutionary meta-analyses (70 – 95%; Senior et al., 2016). High heterogeneity (> 75%) of the effect sizes in our data set indicates variability in the influences of non-pharmacological manipulations of affective state on judgement bias in animals, but is perhaps not surprising given how diverse the studies were in terms of, for example, species used (22 diverse species; Figure 5), task variants, affect manipulations, and other methodological specifics. Accordingly, the lack of phylogenetic effects in our data set is consistent with the observation that meta-analyses on phylogenetically diverse sets of species are unlikely to show a strong phylogenetic signal (Chamberlain et al., 2012).

Key moderators of judgement bias tests

We also revealed five important moderators of responses in the judgement bias task. Four are related to methodology and one is a biological factor. First, active choice go/go tasks tended to yield larger effects than go/no-go tasks. It is possible that the former are more cognitively challenging given that the response needs to be deployed to different stimuli. Such a potential cognitive load might render go/go tasks less susceptible to habitual responding and thereby more sensitive to affect manipulations. Furthermore, go/no-go tasks are likely to be vulnerable to the influence of Pavlovian action predispositions (e.g. go-for-reward; no-go to avoid punishment; Guitart-Masip et al., 2014; Jones et al., 2017), that could inadvertently bias responding (Mendl & Paul, 2020) and obscure affect manipulation effects. Additionally, subjects may sometimes perform no-go responses for reasons unrelated to affect manipulations (e.g. failing to detect or attend to a cue; Bethell, 2015; Jones et al., 2018), making these tests less dependable. Still, we observed that go/no-go tasks are more commonly used in judgement bias studies (in 57 vs. 14 studies; Figure 3), probably because they are easier and quicker to train.

Second, Reward-Reward tasks usually generated larger effect sizes than Reward-Null tasks. Part of the reason for this may be that Reward-Reward tasks usually involve a go/go active choice response and this itself predisposes stronger effects, as just discussed. The most frequently used Reward-Punishment tasks had the largest observed average effect size. It is possible that the Reward-Punishment design, providing a more affectively-laden task (i.e. decision outcomes can range from a desired reward to an aversive punisher), is more sensitive to manipulations of affective state (see Mendl et al. 2009).

Third, the use of auditory and tactile cues tended to reveal the largest effects compared to when spatial, visual and olfactory cues were employed. There may be a number of reasons for this, some of which may be linked to differences in species biology (Bethell, 2015). For example, whilst people are strongly visually focused when information gathering, many other animal

species are not, and may not readily exhibit human-like processing of visual cues. Conversely, olfactory sensitivity in humans is poor, relative to many other species, and this may impair the ability of researchers to design or use meaningful cues in this sensory dimension. It is also possible that cue modality and presentation method can influence the uncertainty of information provided by 'ambiguous' cues. For example, there may be greater uncertainty about the information provided by a single tone intermediate between two training tones, than by a spatial location situated between two training locations. Such differences in uncertainty may have knock-on effects on animal's decisions.

Fourth, cue ambiguity level (P, NP, MID, NN, N) was important. We found predicted judgement bias only at ambiguous cues in the full dataset analysis, and not at positive or negative training cues, on average. Still, some individual studies in the dataset yielded large effects at positive or negative training cues (e.g. Deakin, 2018; Horváth et al., 2016; Zidar et al., 2018). In line with this, Neville et al. (2020) noted that pharmacological manipulations of affect altered judgement bias principally at ambiguous cues, but also at the negative training cue. Large effects at non-ambiguous cues could occur in at least two ways. First, if affect manipulations altered valuation of decision outcomes (e.g. by decreasing food valuation and hence generating a weaker response to the positive cue), the manipulations could change propensities to perform specific responses (e.g. go vs. no-go) and interfere with memory of training cue-outcome associations. Second, large effects at non-ambiguous cues might occur if training was brief or ineffective such that there was considerable ambiguity about the training cue-outcome association during testing (see Mendl et al., 2009; Bateson et al. 2011; Bethell, 2015; Mendl and Paul, 2020). As mentioned earlier, further research of the effects of variation in discrimination training criteria on test performance would shed light on this issue.

Finally, in all analyses, larger predicted effect sizes tended to be reported for male subjects than for females or mixed sex groups. This pattern could be due to existence of sex differences in

neurobiology of learning and memory (Jonasson et al., 2005) or sex differences in stress effects on memory, with different patterns for acute and prolonged stress (Andreano and Cahill, 2009). Effects of enrichment may also be sex-specific (Lin et al., 2011; ter Horst et al., 2012).

Potential limitations and recommendations

The results of our meta-analysis come with six caveats, which we list here alongside recommendations for future studies of judgement bias. First, captive and domesticated mammals dominate the dataset making our conclusions particularly relevant to research on welfare of such animals. Conversely, the analyses are less informative for wild animals, vertebrates other than mammals, and invertebrates. Indeed, Bethell's narrative review (2015) highlighted biased taxonomic representation in empirical evidence. Thus, future work in this area could aim to increase the representation of non-domesticated species, such as those kept in zoos and for research (where animal welfare is of concern; Baumans, 2005; Bethell, 2015; Wolfensohn et al., 2018) and invertebrates (where welfare is an emerging issue; Drinkwater et al., 2019).

Second, we had limited statistical power to detect clear differences between the levels of a number of the tested moderators. Also, the small sample sizes at some levels of the considered moderators might have introduced some spurious findings. For example, relatively few studies used tactile or olfactory cues (e.g. Barker et al., 2017; Novak et al., 2016), and very few used reinforced ambiguous cues during tests (e.g. Bailoo et al., 2018; Keen et al., 2014). To address this limitation, future studies of commonly used laboratory and domesticated species should systematically investigate the role of different cue types. Researchers should also attempt to make cue types relevant for a given species, and vary the perceptual closeness of training cues and hence the difficulty of the task and uncertainty of ambiguous cues.

Third, for some moderators, especially those related to study quality, poor reporting might have obscured statistical relationships. Very few of the included studies explicitly stated that they used automation or blinding, and we had to assume that the remaining studies did not use these. Thus, automation of measurements could be used more often, and/or their use should be clearly reported. Notably, Nguyen et al. (2020) in their systematic review of 20 rodent studies highlighted limited information on the details of experimental procedures and analyses in 65% of assessed studies, undermining confidence in the findings. Nevertheless, we found no statistical evidence for publication bias in our meta-analytic data set. The lack of publication bias is potentially due to our full data set containing data points across the whole response curve, which are usually a mixture of small and large positive effects (in the expected direction) and even some negative ones (not in the expected direction). Also related to reporting, mixed-sex groups of animals comprised almost one-third of the data in our meta-analysis, potentially obscuring sex-specific effects. Providing sex-disaggregated data in research is absolutely essential for improving our understanding of animal behaviour and cognition (Shansky and Woolley, 2016; Palanza and Parmigiani, 2017).

Fourth, we were also not able to include strength of manipulation in our analyses (there is no common scale for the diverse types of manipulations included in our data set). To overcome this problem, in future studies it would be valuable to test and synthesize relationships between measures of cognitive bias and different biomarkers of stress, such as cortisol, adrenaline, alpha-amylase, testosterone, leucocyte profiles (Keay et al., 2006; Davis et al., 2008).

Fifth, we also noted some outliers in the dataset, which usually came from studies with severe manipulations and/or small sample sizes. We, however, conducted extensive sensitivity analyses to test robustness of our conclusions, with the results generally conforming to our predictions and being robust across different statistical approaches. Further, in individual empirical studies comparing two means, to achieve power of 0.8 at alpha of 0.05, it is necessary to have sample

sizes of at least 50 animals per group for detecting moderate effect sizes (Hedges' g = 0.4). Best-case scenario, when effect is large (Hedges' g = 0.8), would require only 13 animals per group to achieve the same power. Conducting power analyses to determine suitable sample sizes for planned experiments can help reducing animal use and also prevent wasting animals on underpowered studies.

Finally, the largest responses often do not appear at the most intermediate/ambiguous cue. Because of this, we suggest that multiple ambiguous (probe) cues (at least 3) are needed for robust and comprehensive judgement bias tests although 25% of response curves in our data set included only one ambiguous cue.

Taken together, while it is unlikely that a single "perfect" version of a judgement bias test exists, our analyses suggest that the most sensitive setup would comprise a go/go active choice task employing a reward-punishment or reward-reward reinforcement contingency, and using at least three ambiguous cues of a sensory modality appropriate to the species of interest. Such tasks may require more lengthy training than conventional go/no-go tasks, but promising new go/no-go variants that achieve many of the benefits of go/go tasks by incorporating active trial initiation are being developed and can be trained relatively rapidly (Hintze et al. 2018; Jones et al. 2018). For more detailed guidance on how tests should be designed and conducted, and what types of adjustments may be needed for different organisms, we refer readers to the works of Bethell (2015) and Roelofs et al. (2016).

Conclusions

In summary, judgement bias tests are a valid method of measuring animal affective state. However, high heterogeneity among studies, which can be only partially explained by simple influences of considered moderators, warrants care in designing and interpreting judgement

bias manipulations and tests. We call for better reporting of experimental designs, especially blinding and automation, disaggregation of data by sex of subjects, and other experimental details that might influence study results. Also, there is a need for more empirical studies that compare different experimental designs and setups, including using different types of tasks, cues, and cue ambiguity levels.

Acknowledgments

We are thankful to all authors that have provided us with additional information or data, as indicated in Table S2. M.L. and S.N. were supported by the Australian Research Council Discovery Project (DP200100367). E.S. and J.Z. were funded by Carl Trygger's Foundation and the Swedish research council Formas, respectively, awarded to H.L.. M.M., E.S.P. and V.N. thank the UK Biotechnology and Biological Sciences Research Council (BBSRC grants BB/P019218/1, BB/T002654/1 and BBSRC SWBio DTP grant BB/M009122/1), and the UK National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs grant NC/K00008X/1) for supporting their work in this area. The authors declare no conflicts of interests.

Data and Code Availability

All data and code are available from the following online OSF repository: https://osf.io/anfhm/

References

Andreano, J.M., Cahill, L., 2009. Sex influences on the neurobiology of learning and memory. Learn. Mem. 16, 248–266. https://doi.org/10.1101/lm.918309

Ash, H., Buchanan-Smith, H.M., 2016. The long-term impact of infant rearing background on the affective state of adult common marmosets (*Callithrix jacchus*). Appl. Anim. Behav. Sci. 174, 128–136. https://doi.org /10.1016/j.applanim.2015.10.009

1 Baciadonna, L., McElligott, A.G., 2015. The use of judgement bias to assess welfare in farm
2 livestock. Anim. Welf. 24, 81–91. https://doi.org/10.7120/09627286.24.1.081

3 Bailoo, J.D., Murphy, E., Boada-Saña, M., Varholick, J.A., Hintze, S., Baussière, C., Hahn, K.C.,
4 Göpfert, C., Palme, R., Voelkl, B., Würbel, H., 2018. Effects of cage enrichment on behavior,
5 welfare and outcome variability in female mice. Front. Behav. Neurosci. 12, 232.
6 https://doi.org/10.3389/fnbeh.2018.00232.

7 Barker, T.H., Bobrovskaya, L., Howarth, G.S., Whittaker, A.L., 2017. Female rats display fewer
8 optimistic responses in a judgment bias test in the absence of a physiological stress
9 response. Physiol. Behav. 173, 124–131. https://doi.org/10.1016/j.physbeh.2017.02.006

10 Barton, K., 2009. MuMIn: Multi-model inference. R Package Version 0.12.2/r18. http://R-
11 Forge.R-project.org/projects/mumin/

12 Bateson, M., 2016. Optimistic and pessimistic biases: a primer for behavioural ecologists. Curr.
13 Opin. Behav. Sci. 12, 115–121. https://doi.org/10.1016/j.cobeha.2016.09.013

14 Bateson, M., Desire, S., Gartside, S. E., Wright, G. A., 2011. Agitated honeybees exhibit pessimistic
15 cognitive biases. Curr. Biol. 21, 1070–1073. doi: 10.1016/j.cub.2011.05.017

16 Baumans, V., 2005. Science-based assessment of animal welfare: laboratory animals. Rev. Sci.
17 Tech. 24, 503-513. http://dx.doi.org/10.20506/rst.24.2.1585

18 Bethell, E.J., 2015. A "how-to" guide for designing judgment bias studies to assess captive animal
19 welfare. J. Appl. Anim. Welf. Sci. 18, S18–S42. https://doi.org/
20 10.1080/10888705.2015.1075833.

21 Bethell, E.J., Holmes, A., MacLarnon, A., Semple, S., 2016. Emotion evaluation and response
22 slowing in a non-human primate: new directions for cognitive bias measures of animal
23 emotion? Behav. Sci. (Basel). 6, 2. https://doi.org/10.3390/bs6010002

24 Bethell, E.J., Koyama, N.F., 2015. Happy hamsters? Enrichment induces positive judgement bias
25 for mildly (but not truly) ambiguous cues to reward and punishment in *Mesocricetus*
26 *auratus*. R. Soc. Open Sci. 2, 140399. https://doi.org/10.1098/rsos.140399

27 Blanchette, I., Richards, A., 2010. The influence of affect on higher level cognition: a review of
28 research on interpretation, judgement, decision making and reasoning. Cogn. Emot.
29 https://doi.org/10.1080/02699930903132496

30 Brilot, B.O., Asher, L., Bateson, M., 2010. Stereotyping starlings are more "pessimistic." Anim.
31 Cogn. 13, 721–731. https://doi.org/10.1007/s10071-010-0323-z

Burman, O.H.P., Mendl, M.T., 2018. A novel task to assess mood congruent memory bias in non-human animals. J. Neurosci. Methods 308, 269–275. https://doi.org/10.1016/j.jneumeth.2018.07.003

Burman, O.H.P., Parker, R.M.A., Paul, E.S., Mendl, M.T., 2009. Anxiety-induced cognitive bias in non-human animals. Physiol. Behav. 98, 345–350. https://doi.org/10.1016/j.physbeh.2009.06.012

Chamberlain, S.A., Hovick, S.M., Dibble, C.J., Rasmussen, N.L., Van Allen, B.G., Maitner, B.S., Ahern, J.R., Bell‐Dereske, L.P., Roy, C.L., Meza‐Lopez, M., Carrillo, J., Siemann, E., Lajeunesse, M.J. and Whitney, K.D., 2012. Does phylogeny matter? Assessing the impact of phylogenetic information in ecological meta‐analysis. Ecol. Lett. 15, 627-636. doi:10.1111/j.1461-0248.2012.01776.x

Crump, A., Arnott, G., Bethell, E.J., 2018. Affect-Driven Attention Biases as Animal Welfare Indicators: Review and Methods. Animals 8. https://doi.org/10.3390/ani8080136

da Cunha Nogueira, S.S., Fernandes, I.K., Oliveira Costa, T.S., Gama Nogueira-Filho, S.L., 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

Cohen, J. (1969). Statistical power analysis for the behavioral sciences (1st ed.). New York, NY: Academic Press.

Darwin, C., 1872. The expression of the emotions in man and animals, 3rd ed., The expression of the emotions in man and animals, 3rd ed. Oxford University Press, New York, NY, US.

Davis, A.K., Maney, D.L., Maerz, J.C., 2008. The use of leukocyte profiles to measure stress in vertebrates: a review for ecologists. Funct. Ecol. 22: 760-772. https://doi.org/10.1111/j.1365-2435.2008.01467.x

Destrez, A., Deiss, V., Lévy, F., Calandreau, L., Lee, C., Chaillou-Sagon, E., Boissy, A., 2013. Chronic stress induces pessimistic-like judgment and learning deficits in sheep. Appl. Anim. Behav. Sci. 148, 28–36. https://doi.org/10.1016/j.applanim.2013.07.016

Douglas, C., Bateson, M., Walsh, C., Bédué, A., Edwards, S.A., 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

Doyle, R.E., Lee, C., Deiss, V., Fisher, A.D., Hinch, G.N., Boissy, A., 2011. Measuring judgement bias and emotional reactivity in sheep following long-term exposure to unpredictable and

aversive events. Physiol. Behav. 102, 503–510. https://doi.org/10.1016/j.physbeh.2011.01.001

Drinkwater, E., Robinson, E.J.H., Hart, A.G., 2019. Keeping invertebrate research ethical in a landscape of shifting public opinion. Methods Ecol Evol. 10, 1265–1273. https://doi.org/10.1111/2041-210X.13208

Egger, M., Smith, G.D., Schneider, M., Minder, C., 1997. Bias in meta-analysis detected by a simple , graphical test measures of funnel plot asymmetry. BMJ 315, 629–634. https://doi.org/10.1136/bmj.315.7109.629

Girard, B., Bellone, C., 2020. Revealing animal emotions. Science 368, 33–34. doi: 10.1126/science.abb2796

Guitart-Masip M., Duzel, E., Dolan, R., Dayan, P., 2014. Action versus valence in decision making. Trends Cogn. Sci. 18,194-202. https://doi.org/10.1016/j.tics.2014.01.003

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

Hadfield, J.D., 2010. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. J. Stat. Softw. 33, 1–22. https://doi.org/10.18637/jss.v033.i02

Hales, C.A., Robinson, E.J., Houghton, C.J., 2016. Diffusion modelling reveals the decision making processes underlying negative judgement bias in rats. PLOS One 11. https://doi.org/10.1371/journal.pone.0152592

Hales, C.A., Stuart, S.A., Anderson, M.H., Robinson, E.S., 2014. Modelling cognitive affective biases in major depressive disorder using rodents. Br. J. Pharmacol. 171, 4524–4538. https://doi.org/10.1111/bph.12603

Harding, E.J., Paul, E.S., Mendl, M., 2004. Cognitive bias and affective state. Nature 427, 312.

Hedges, L. V., Olkin, I., 1985. Statistical Methods for Meta-Analysis, Statistical Methods for Meta-Analysis. New York: Academic Press. https://doi.org/10.1016/c2009-0-03396-0

Hintze, S., Melotti, L., Colosio, S., Bailoo, J.D., Boada-Sana, M., Wurbel, H., Murphy, E., 2018. A cross-species judgement bias task: integrating active trial initiation into a spatial Go/No-go task. Sci. Rep. 8, 5104. doi:10.1038/s41598-018-23459-3.

Jennions, M.D., Møller, A.P., 2002. Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. Proceedings. Biol. Sci. 269, 43–8. https://doi.org/10.1098/rspb.2001.1832

Jonasson, Z., 2005. Meta-analysis of sex differences in rodent models of learning and memory: a review of behavioral and biological data. Neurosci. Biobehav. Rev. https://doi.org/10.1016/j.neubiorev.2004.10.006

Jones, S., Neville, V., Higgs, L., Paul, E.S., Dayan, P., Robinson, E.S.J., Mendl, M., 2018. Assessing animal affect: an automated and self-initiated judgement bias task based on natural investigative behaviour. Scientific Reports 8, 12400. doi: 10.1038/s41598-018-30571-x

Jones, S., Paul, E.S., Dayan, P., Robinson, E.S.J., Mendl, M., 2017. Pavlovian influences on learning differ between rats and mice in a counter-balanced Go/NoGo judgement bias task. Behavioural Brain Research 331:214-224. doi: 10.1016/j.bbr.2017.05.044

Keay, J.M., Singh, J., Gaunt, M.C., Kaur, T., 2006. Fecal glucocorticoids and their metabolites as indicators of stress in various mammalian species: a literature review. J. Zoo Wildlife Med., 37, 234-244. https://doi.org/10.1638/05-050.1

Keen, H.A., Nelson, O.L., Robbins, C.T., Evans, M., Shepherdson, D.J., Newberry, R.C., 2014. Validation of a novel cognitive bias task based on difference in quantity of reinforcement for assessing environmental enrichment. Anim. Cogn. 17, 529–541. https://doi.org/10.1007/s10071-013-0684-1

Lin, E.J., Choi, E., Liu, X., Martin, A., & During, M.J. (2011). Environmental enrichment exerts sex-specific effects on emotionality in C57BL/6J mice. Behav. Brain Res. 216, 349–357. https://doi.org/10.1016/j.bbr.2010.08.019

Mendl, M., Burman, O.H.P., Parker, R.M.A., Paul, E.S., 2009. Cognitive bias as an indicator of animal emotion and welfare: emerging evidence and underlying mechanisms. Appl. Anim. Behav. Sci. 118, 161–181. https://doi.org/10.1016/j.applanim.2009.02.023

Mendl, M., Burman, O.H.P., Paul, E.S., 2010. An integrative and functional framework for the study of animal emotion and mood. Proc. R. Soc. B Biol. Sci. 277, 2895–2904. https://doi.org/10.1098/rspb.2010.0303

Mendl, M., Paul, E.S. 2020. Animal affect and decision-making. Neurosci. Biobehav. Rev. 112, 144-163. https://doi.org/10.1016/j.neubiorev.2020.01.025

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Altman, D., Antes, G., Atkins, D., Barbour, V., Barrowman, N., Berlin, J.A., Clark, J., Clarke, M., Cook, D., D'Amico, R., Deeks, J.J., Devereaux, P.J., Dickersin, K., Egger, M., Ernst, E., Gøtzsche, P.C., Grimshaw, J., Guyatt, G., Higgins, J., Ioannidis, J.P.A., Kleijnen, J., Lang, T., Magrini, N., McNamee, D., Moja, L., Mulrow, C., Napoli, M., Oxman, A., Pham, B., Rennie, D., Sampson, M., Schulz, K.F., Shekelle, P.G., Tovey, D.,

Tugwell, P., 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. https://doi.org/10.1371/journal.pmed.1000097

Murphy, E., Nordquist, R.E., van der Staay, F.J., 2013. Responses of conventional pigs and Gottingen miniature pigs in an active choice judgement bias task. Appl. Anim. Behav. Sci. 148, 64–76. https://doi.org/10.1016/j.applanim.2013.07.011

Nakagawa, S., Johnson, P.C.D., Schielzeth, H., 2017. The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. J. R. Soc. Interface 14. https://doi.org/10.1098/rsif.2017.0213

Nakagawa, S., Lagisz, M., O'Dea, R.E., Rutkowska, J., Yang, Y., Noble, D., Senior, A.M., 2020. The orchard plot: cultivating forest plots for use in ecology, evolution and beyond. Res. Syn. Meth. 1– 9. https://doi.org/10.1002/jrsm.1424

Nakagawa, S., Santos, E.S.A., 2012. Methodological issues and advances in biological meta-analysis. Evol. Ecol. https://doi.org/10.1007/s10682-012-9555-5

Nakagawa, S., Schielzeth, H., 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. Meth. Ecolol. Evol. 4, 133–142 https://doi.org/10.1111/j.2041-210x.2012.00261.x

Neville, V., Nakagawa, S., Zidar, J., Paul, E.S., 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

Nguyen, H.A.T., Guo, C., Homberg, J.R., 2020. Cognitive bias under adverse and rewarding conditions: a systematic review of rodent studies. Front. Behav. Neurosci. 14:14. doi: 10.3389/fnbeh.2020.00014

Novak, J., Stojanovski, K., Melotti, L., Reichlin, T.S., Palme, R., Würbel, H., 2016. Effects of stereotypic behaviour and chronic mild stress on judgement bias in laboratory mice. Appl. Anim. Behav. Sci. 174, 162–172. https://doi.org/10.1016/j.applanim.2015.10.004

Oliveira, F.R.M., Nogueira, S.L.G., Sousa, M.B.C., Dias, C.T.S., Mendl, M., Nogueira, S.S.C., 2016. Measurement of cognitive bias and cortisol levels to evaluate the effects of space restriction on captive collared peccary (Mammalia, Tayassuidae). Appl. Anim. Behav. Sci. 181, 76–82. https://doi.org/10.1016/j.applanim.2016.05.021

Palanza, P., Parmigiani, S., 2017. How does sex matter? Behavior, stress and animal models of neurobehavioral disorders. 76A: 134-143. Neurosci. Biobehav. Rev.

https://doi.org/10.1016/j.neubiorev.2017.01.037

Paul, E.S., Harding, E.J., Mendl, M., 2005. Measuring emotional processes in animals: the utility of a cognitive approach. Neurosci. Biobehav. Rev. 29, 469–491. https://doi.org/10.1016/j.neubiorev.2005.01.002

Paul, E.S., Sher, S., Tamietto, M., Winkielman, P., Mendl, M.T., 2020. Towards a comparative science of emotion: affect and consciousness in humans and animals. Neurosci. Biobehav. Rev. https://doi.org/10.1016/j.neubiorev.2019.11.014

R Development Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Roelofs, S., Boleij, H., Nordquist, R.E., van der Staay, F.J., 2016. Making decisions under ambiguity: Judgment bias tasks for assessing emotional state in animals. Front. Behav. Neurosci. 10.

Rolls, E.T., 2009. Emotion Explained. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198570035.001.0001

Rosenthal, R., 1979. The file drawer problem and tolerance for null results. Psychol. Bull. 86, 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rygula, R., Papciak, J., Popik, P., 2013. Trait pessimism predicts vulnerability to stress-induced anhedonia in rats. Neuropsychopharmacology 38, 2188–2196. https://doi.org/10.1038/npp.2013.116

Scherer, K.R., 2005. What are emotions? And how can they be measured? Soc. Sci. Inf. 44, 695–729. https://doi.org/10.1177/0539018405058216

Senior, A.M., Grueber, C.E., Kamiya, T., Lagisz, M., O'Dwyer, K., Santos, E.S.A., Nakagawa, S., 2016. Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and implications. Ecology 97. https://doi.org/10.1002/ecy.1591

Shansky, R.M., Woolley, C.S., 2016. Considering sex as a biological variable will be valuable for neuroscience research. J. Neurosci. 36, 11817–11822. https://doi.org/10.1523/JNEUROSCI.1390-16.2016

ter Horst, J.P., de Kloet, E.R., Schächinger, H., & Oitzl, M S. (2012). Relevance of stress and female sex hormones for emotion and cognition. Cell. Mol. Neurobiol. 32, 725–735. https://doi.org/10.1007/s10571-011-9774-2

Viechtbauer, W., 2010. Conducting meta-analyses in R with the metafor package. J. Stat. Softw. https://doi.org/10.1103/PhysRevB.91.121108

Wallace, B.C., Small, K., Brodley, C.E., Lau, J., Trikalinos, T.A., 2012. Deploying an interactive machine learning system in an Evidence-based Practice Center: Abstrackr, in: IHI'12 - Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. pp. 819–823. https://doi.org/10.1145/2110363.2110464

Wheeler, R.R., Swan, M.P., Hickman, D.L., 2015. Effect of multilevel laboratory rat caging system on the well-being of the singly-housed Sprague Dawley rat. Lab. Anim. 49, 10–19. https://doi.org/10.1177/0023677214547404

Wolfensohn, S., Shotton, J., Bowley, H., Davies, S., Thompson, S., Justice, W.S.M., 2018. Assessment of welfare in Zoo animals: towards optimum quality of life. Animals 8, 110. https://doi.org/ 10.3390/ani8070110

Zidar, J., Campderrich, I., Jansson, E., Wichman, A., Winberg, S., Keeling, L., Løvlie, H., 2018. Environmental complexity buffers against stress-induced negative judgement bias in female chickens. Sci. Rep. 8, 5404. https://doi.org/10.1038/s41598-018-23545-6

Figure captions

Figure 1

Conceptual diagram presenting the main elements of a typical judgement bias study. a) The basic task is trained using either a go/no-go, or active choice (go/go) design. b) Manipulations of affective state usually, but not always, occur after training of the task and may be acute or longer-term. c) Tests involve the standard training protocol plus the addition of occasionally presented ambiguous cues whose properties are usually intermediate between the trained positive and negative cues (NP = near positive cue, MID = intermediate between positive and negative cue, NN = near negative cue). d) 'Optimistic' and 'pessimistic' responding to the cues is inferred from the proportion of positive responses and/or the latency to make positive responses, which are inversely related.

Figure 2

PRISMA flow diagram. Articles identified and number of articles included and excluded during each screening stage.

Figure 3

Main characteristics of the included studies. Blue bars represent numbers of studies represented in each level of categorical variables. Between one and 30 effect sizes were extracted per study and the distribution of effect sizes generally follows the pattern of the presented data aggregated to the study level (e.g. some studies reported data for only one sex, others reported data for both sexes together, and 3 studies that included both sexes reported data for females and males separately, shown here as 'female and male sep.'). Numbers do not add up to 71 for some of the variables due to multiple experiments being present within some studies, or complex experimental designs being used.

Figure 4

Forest-like (orchard) plots showing effect size (Hedges' g) estimates from meta-analyses on: a) whole data set (all cues reported for judgement bias tests), and b-e) four subsets of this data set, representing different ways of interpreting the judgement bias test results. Positive effect sizes

indicate a positive effect of affect manipulation treatments on judgement bias in a relatively better condition compared to a relatively worse condition, i.e. affect manipulations working in the expected direction. The effects are statistically significant when the thick horizontal error bars (95% confidence intervals) do not cross zero. Thin horizontal whiskers indicate prediction intervals. *k* is number of effect sizes. Dots represent individual effect sizes scaled proportionally to their precision.

Figure 5

Forest plot showing mean effect size (Hedges' *g*) estimates from meta-regression analysis using species identity as a moderator. Positive effect sizes indicate a positive effect of affect manipulation treatments on judgment bias in a relatively better condition compared to a relatively worse condition, i.e. affect manipulations working in the expected direction. The effects are statistically significant when the horizontal error bars (95% confidence intervals) do not cross zero. *k* is number of effect sizes, *K* is number of studies.

Figure 6

Forest plots showing effect size (Hedges' *g*) estimates from the univariate meta-regression analyses (one moderator at a time) with potentially influential moderators. Effect sizes with positive values indicate a positive effect of affect manipulations on judgement bias in a relatively better condition compared to a relatively worse condition, i.e. affect manipulation treatment working in the expected direction. The mean effects (black unfilled circles) for each group of individual effect sizes (grey filled circles) are statistically different from zero when their horizontal error bars (95% confidence intervals) do not cross zero. Thin horizontal whiskers indicate prediction intervals. *k* is number of effect sizes. Dots represent individual effect sizes scaled proportionally to their precision.