Biology & Philosophy (2022) 37: 36
https://doi.org/10.1007/s10539-022-09862-1

Assessing measures of animal welfare

Heather Browning[1]

Received: 5 December 2021 / Accepted: 9 June 2022 / Published online: 12 August 2022

Abstract
There are many decision contexts in which we require accurate information on animal welfare, in ethics, management, and policy. Unfortunately, many of the methods currently used for estimating animal welfare in these contexts are subjective and unreliable, and thus unlikely to be accurate. In this paper, I look at how we might apply principled methods from animal welfare science to arrive at more accurate scores, which will then help us in making the best decisions for animals. I construct and apply a framework of desiderata for welfare measures, to assess the best of the currently available methods and argue that a combined use of both a whole-animal measure and a combination measurement framework for assessing welfare will give us the most accurate answers to guide our action.

Keywords Animal welfare · Measurement · Policy · Whole-animal measure · Combination measurement framework

Introduction

Animal welfare is important in a large range of contexts. Most people agree that animal welfare is morally important: it is bad for animals to suffer and good for them to have happy lives, and we should act where possible to prevent the former and enable the latter. Animal welfare can be taken to mean different things – as with human wellbeing, there are theories of animal welfare that take welfare to consist in different subjective or objective goods (Browning 2020; Veit and Browning 2021a). Here, I take a subjective, or hedonic, view of animal welfare, in which welfare consists in the subjective mental states experienced by an animal - "the quality of its emotional

✉ Heather Browning
drheatherbrowning@gmail.com

[1] London School of Economics and Political Science, Centre for Philosophy of Natural and Social Science, London, UK

states, including their sign (positive or negative), intensity and duration" (Bracke 2001, p. 45). This concept has been chosen for two primary reasons, that I will briefly summarise here (see Browning 2020 for a more extended defence of the use of a subjective welfare concept). The first is that subjective experiencing is morally significant – it is almost universal to take the capacity to suffer as a morally relevant feature when deciding how to treat other animals. The second is that subjective experience is biologically relevant, giving us a perspective on welfare from the point of view of the animal itself, and what is in its interests, what it likes, wants, and needs.

Subjective animal welfare is a common view, particularly within animal welfare science (e.g. Duncan 2002; Mellor et al. 2020), and the dominant view within some of the contexts discussed, such as within effective altruism. Even where one rejects this as a complete view of welfare, it is still true that the positive and negative experiences of animals – their pleasure and suffering – comprise at least part of animal welfare under any conception, and thus are important to measure. Thus where one holds a different conception of animal welfare, the framework I present here can still be used to assess measures of welfare, though the conclusions about their relative usefulness would likely be different. For any decisions aimed at decreasing suffering and/or increasing pleasure for animals, this work will be relevant. Indeed, many of our decisions can have impacts on animals, and their interests are a source of value that should count in this decision-making.

In particular, there are many decision contexts in which we require accurate information regarding the welfare of animals under different conditions, in order to evaluate and compare the costs and benefits of different possible actions. These include policy deliberations by governments and businesses, and prioritising charitable giving and interventions, on both an individual and organisational level. Making these decisions requires methods of accurately measuring the welfare status of different animals to perform the necessary calculations, and for this reason the question of measuring subjective animal welfare is not just scientific but one that is relevant also to philosophers and policy-makers. My aim in this paper is to critically assess some of the best available methods of performing such measures.

Policy analyses assess the value of different outcomes of some policy decision, typically based on an impact assessment judging the impacts on wellbeing (or some proxy), as well as a social welfare function that aggregates these values into an overall societal value of the outcome, and cost-benefit analyses that can compare the relative value of different outcomes relative to their costs (Budolfson and Spears 2020b). This then raises concerns about the justification for aggregation of the welfare of different individuals into a single score, trading off the sufferings of some against the pleasures of others, and the construction of an adequate social welfare function must take this into account. However, for the purposes of this paper, I will assume that insofar as we are or have been able to do so for humans, we can apply similar methods for the case of animals (so too for the charitable giving contexts described shortly).

Often, these policy assessments include animal welfare only indirectly, in terms of human preferences for improved animal welfare. However, there have been recent calls for the direct inclusion of animal welfare alongside humans within these analyses, for a more complete and accurate picture of the welfare costs and benefits of different decisions (Budolfson and Spears 2020a, b; Carlier and Treich 2020). One

example is in assessments of the negative externalities of some market transactions (i.e. harms to individuals or society that are not internalised in market prices), such as including the welfare costs to the individual animals from animal agriculture, within the pricing structures for agricultural products (Kuruc and McFadden 2021; Lusk and Norwood 2011). Another is in climate change policies, as many of the individuals affected by climate change will be animals, both domestic and wild (Budolfson and Spears 2020b; Sunstein and Hsiung 2006). It also includes more generally impact assessments on the effects of changes in policy, or land-use and developments that have the potential to affect animals (Sebo 2022); an example of which can be seen in the argument for the use of the Animal Welfare Impact Assessment for decisions that affect sentient animals, such as badger culling (discussed in McCulloch & Reiss, 2017).

Though this current lack of inclusion is in large part a result of anthropocentric biases, it is also in part because policymakers don't necessarily have good ways of quantifying impacts on animal welfare in the same ways as they do for humans, where economics has developed a range of appropriate proxies. Budolfson and Spears (2020a, 2020b) have identified two components to this problem – that of identifying the welfare experience of individual animals under different conditions, and that of finding a way of weighting this against impacts on human wellbeing. The latter, also known as the problem of interspecies comparisons, is a complex one I have in part addressed elsewhere (Browning 2020); probably requiring the use of proxies for welfare capacity, or setting conventions regarding moral weights. In this paper, I am interested in the first part of the problem, of providing methods to quantify subjective animal welfare, or quality of life, as an input into these calculations.

Another context where accurate animal welfare measures are important is in prioritising charitable giving and interventions. Under conditions of scarce resources, decisions about investment or action will depend on where one can have the most impact. The number of organisations aimed at determining and undertaking the most effective charitable actions is growing, including many with a particular focus on animal welfare improvements (e.g. Animal Charity Evaluators, Animal Ask, Animal Ethics, Faunalytics, Wild Animal Initiative). Typically here, the desired impact is welfare gains and thus calculations require knowledge of the welfare gains of different options, based in an accurate understanding of the quality of life of animals living under different conditions. Accurate measures of animal welfare are required to identify and enact the most relevant and effective actions.

Within this sphere, there have been attempts to create 'suffering calculators' or similar welfare estimation frameworks that aim to compare the total suffering produced by different types of production systems in order to determine where resources would be best invested. While some of this work is published (e.g. Alonso and Schuck-Paim 2021; Norwood & Lusk, 2011; Scherer et al. 2018), much can also only be found online (e.g. Charity Entrepreneurship[1], Essays on Reducing suffering [Tomasik 2018], Warren 2018). Welfare calculators are created to compare the total suffering produced by different sets of conditions – most often different agricultural

[1] https://www.charityentrepreneurship.com/blog/is-it-better-to-be-a-wild-rat-or-a-factory-farmed-cow-a-systematic-method-for-comparing-animal-welfare.

systems. These calculators take a variety of different types of information on the numbers of animals used, the length of life of these animals and the quality of their life (or amount of suffering experienced) on an average day, sometimes also taking into account the impact of rare or unusual experiences, such as veterinary procedures, handling and slaughter. Additional inputs are often used, such as a 'weighting factor' (referring to the relative welfare capacity or moral weight for a species) and a 'badness of death' measure (quantifying how bad the loss of life is for an animal). From these, an overall calculation can be performed of the comparative impact of different systems. For instance, the calculator produced by Tomasik shows that using his estimates, catfish farms produce over 10,000x more suffering than dairy farms. This would then give us impetus to act to reduce the suffering of farmed catfish, either by reducing their numbers, or improving their lives.

A number of other websites and charity evaluators use similar calculators to measure the number of equivalent years of suffering that can be saved per dollar donated, often for specific species only (e.g. Open Philanthropy Project[2], Animal Charity Evaluators[3]). As most of the animals in human care are livestock animals used for food production, this has usually been the area of focus for research of this type, though there is increasingly work on wild animals (Harvey et al. 2020; e.g. Ng 2016; Tomasik 2015; Veit and Browning 2021b), including proposals for developing the new field of 'welfare biology' which uses the tools of ecology to expand animal welfare science to encompass the complexities of assessing interventions in wild animal populations (Faria and Horta 2019; Soryl et al. 2021). Here, I am interested in how we can ensure we are inputting the right information into such calculators to get accurate outputs of comparative suffering or overall welfare, with which to guide our decisions.

The problem with the welfare score as it has been used so far, is that this measure is computed in a number of vastly different ways, which can then lead to vastly different results. Table 1 (adapted from Warren 2018) shows the estimates given by a few of the more commonly used models, and demonstrates how much they differ. In some cases, the sign of the score is different, indicating that under some measures it comes out as positive (a life of mainly positive experiences; a life worth living) and others it comes out negative (a life of mainly negative experiences; not worth living). Look, for instance at the range of scores given to turkeys: from −57 to +3. This is clearly a problem if these scores are a critical part of the calculations that are supposed to guide our decision-making. We would end up endorsing what could be quite different courses of action, depending on which estimate we chose to use.

There is also an additional issue regarding the differences between different instances of a single system type. Even while there are many similarities between different production facilities of the same type, there are also differences that can greatly impact welfare, such as the skills and knowledge of the animal managers and handlers (Fraser 2014). Thus we should also be cautious about extrapolating from the information about any one context, even to other systems of the same type, without reflecting on the specifics of the relevant similarities and differences. However, a sur-

[2] https://www.openphilanthropy.org/blog/initial-grants-support-corporate-cage-free-reforms.

[3] https://animalcharityevaluators.org/research/methodology/our-use-of-cost-effectiveness-estimates/.

Table 1 Suffering estimates. (adapted from Warren 2018)

Farmed Animals	Warren	Norwood	Shields	Norowitz	Tomasik	Scherer et al.	Charity Ent.
Beef	+6	+6	+2	+6	1 (reference)	0.66	-20
Dairy	0	+4	0	-4	x2	0.76	-34
Fish	-5		-7	-7	x 1.5	1.0	-44
Pork	-5	-2	-5	-10	x 2.5	0.80	
Turkeys	-6	+3	-8	-11	x3	0.39	-57
Broilers	-6	+3	-8	-13	x3	0.39	-56
Cage-Free Hens	-7	+2		-7			
Veal	-7	-8					
Caged-Hens	-8	-8	-7	-25	x4	0.60	-57

vey of a range of individual producers within a system may help offset this to some degree, by providing an average score, or range of scores, that can be used to describe a particular type of production system and how it typically compares to others.

We have a range of decision contexts that require an accurate measure of animal welfare, with no currently established principled way of filling in these values. Here, I suggest that decision-makers involved in policy or prioritising charitable interventions should look to animal welfare science for appropriate methods. There are, of course, potential issues with animal welfare science as a discipline. Several scholars have raised concerns about animal welfare science as it is currently practiced (Bekoff and Pierce 2017; Cooke 2021; Haynes 2008; Pierce 2019). In particular, that it is used primarily *within* animal industries and could serve to further their interests, rather than being independently focussed on the interests of animals. Within the context of this paper, I do not see this as being too problematic, for two reasons. The first is that here I am merely interested with the features of the specific measurement tools used within animal welfare science, rather than the contexts in which they are typically employed, which is where the problems seem to arise. One can concede that the tools are valid, even while maintaining that they are often used to further the wrong ends. The second is that animal welfare science is currently the only good source of methods for the quantification of animal welfare; even while one might hope that this changes in future. It is therefore possible to assess the best currently available methods of measurement, while leaving open the possibility (and even desirability) of developing new ones that better fit the goals these scholars propose.

This may be seen as an instance of the more general discussion of the role of values within animal welfare science (see e.g. Fraser 2008; Lassen et al. 2006; Sandøe et al. 2003). Animal welfare is as much a moral concept as it is a scientific concept, and thus different values will come into play at various stages within the measurement of animal welfare – from the choice of welfare concept (as discussed above) to the selection of indicators to measure welfare and the weighting of different factors contributing to welfare, and in decisions about which actions to take based on the results of assessments. This is important to keep in mind whenever one is assessing the measures of animal welfare, and in the discussion in Sects. Whole-Animal Measures and Combination measures, I will indicate where this may most strongly influence the measures used.

As this illustrates, not all measures coming out of welfare science will be equally fit for the required purposes and so will need to be critically assessed for their use. In this paper, I will begin in Sect. Desiderata for a welfare index by constructing a framework of desiderata for a good measure of animal welfare relevant to this purpose, grouped into the categories of correctness, usefulness and feasibility. I will then go on in Sects. Whole-Animal Measures and Combination measures to assess a range of possible candidate measurement methods according to these desiderata, with recommendations as to which are likely to give us the best results. I will finish in Sect. Conclusion by looking at the upshots of these considerations and identifying some useful areas for future work.

Desiderata for a welfare index

When trying to decide which is the best measure to use in quantifying animal welfare for the purposes of political and ethical decision-making, we need to have in mind what features this measure must have to make it fit for this purpose. In this section I will develop a framework for assessing potential measures, grouping the criteria into three categories – correctness, usefulness, and feasibility. While some of these are general criteria for the quality of almost any measure, others are more specific to subjective animal welfare and the decision-making contexts discussed. I will then move on in the sections that follow to look at how well different methods of measuring welfare meet these criteria.

Correctness criteria

The first set of criteria are the correctness criteria, which represent the degree to which the measure will give the right results to use in relevant calculations – that is, numbers that really do reflect the welfare as experienced by the animals. These are the most crucial criteria, as without the right inputs, any results generated will be meaningless.

Validity

The first criterion, and probably the most important, is validity. A measure is valid if it is measuring the intended target, instead of some other property or state. The reason this is central is that if a measure is not valid – if it is not actually measuring animal welfare - it does not matter how well it meets the other criteria. It is thus important to be very clear about the target state – the integrated set of mental states that constitute welfare – to ensure the measure is tracking this and only this. While different conceptions of welfare may include other components of welfare, these are not the targets for this project. It is not sufficient to have a broad category of those things which matter to us ethically with regards to animals, only those relating to welfare as experienced by the animal. Otherwise a measure could produce misleading results, and lead to recommendations of actions not actually beneficial to animals. Taking a

pre-defined notion of welfare and then assessing validity relative to this is a better way of ensuring we hit our intended target.

Validity can be tested through the presence of reliable correlations between changes in the measure and changes in the target state, particularly under experimental manipulations, as this helps rule out non-causal correlations that would undermine validity. For subjective animal welfare, where the target state (subjective experience) is hidden from direct measurement, this can still be achieved through correlations with other established measures, or through using manipulations in upstream variables (such as husbandry inputs) to create changes in downstream variables (such as animal-based measurement indicators) to establish causal connections (see Browning 2020 for details). Measures can thus be assessed on whether and how well they have performed this validation process.

Accuracy

As well as being valid (measuring the intended target), the measure should be accurate. This means that the measured values are close to the actual values in the target system – that when subjective welfare is high, the measured values are high, and the same for medium, low, neutral etc. It also includes sensitivity in detecting relevant changes in welfare: i.e., when there are small increases or decreases in an animal's welfare experience, the measured values will change accordingly. Particularly in cases where we are comparing quite similar systems or looking at the impact of different interventions on a system, while the individual changes might be quite small, the total impact could still be large if a large number of animals are affected. Insensitive measures that fail to track such changes will not provide the right recommendations.

It is possible for a measure to be valid, and measuring the correct target, but still inaccurate because it does so poorly. For example, think of making estimates of environmental temperature based on one's subjective 'feeling' of how hot or cold it is. I might make a guess that the outdoor temperature is in the low 20s, based on how warm I feel. This is a valid measure, as I am responding to environmental temperature, and not some other state. However, it is measure with low accuracy, as I am likely to have the value correct only within a range of around ±5 °C. It would also be possible to have a measure which is accurate, but not valid, as it is not measuring the intended target, but some other target - perhaps a common cause which creates changes both in the target variable and the measure.

Completeness

A measure of animal welfare intended for the decision contexts I have described has to be complete, providing a comprehensive assessment of the entire state of subjective welfare of the animal. A measure that only represents some part of the animal's experience, leaving out or overlooking some aspects, will fail for this purpose. The measure should incorporate the different affects that make up subjective welfare experience, or all the conditions that contribute to it. For example, some measures may reflect only physical health, while not accounting for psychological contributors to welfare, but these will then not provide an accurate score, leading to wrong recom-

mendations. In a sense, this is part of accuracy, as an incomplete measure will give inaccurate results, but as there are many welfare measures that vary in their degree of completeness, it is worth drawing attention to and assessing independently.

Reliability

The final correctness criterion is reliability, meaning the measurement method should give consistent results when repeated, with low variation between repeated measures. Repetition in this sense can be of many kinds (Czycholl et al., 2015), and ideally our measure should be reliable across all of them, including intra-observer (multiple repeated measures taken by the same observer), inter-observer (measures taken by different observers, of the same target), and test-retest (results produced at different times and under different conditions). Where reliability is low, this reduces the likelihood that the results produced by any particular test were accurate, or even that the test is valid.

Usefulness criteria

The correctness criteria described above are the most significant for selecting the right measure, as they ensure that the results produced are the right ones. However, it is also important that the outputs of the measure will do well for the task required. Usefulness criteria describe how well the outputs of the measures fill the role we require them for in providing useful data for the contexts previously described.

Range of applicability

As discussed, there are a range of different contexts that require quantified animal welfare inputs to aid decision-making. Ideally, a measure should be useful across the full range of contexts of interest. Perhaps most importantly, this means using the same measure for all species being investigated, as using different measures for different species risks weakening the comparisons. There is a large range of animal species these decisions will cover - from large mammals through to the insects and shrimp now used in farming systems - and the measure should be applicable to all of them. This still leaves open the issue of how to standardise scores to make interspecies comparisons – no measure will be able to both produce a welfare score for a species and indicate how to scale it appropriately – but as discussed earlier, this is a more complex issue that needs to be dealt with independently.

It should also be applicable across the different types of animal usage, from livestock to wild animals, to increase the scope of decision-making power. A measure that is useful only in a small range of circumstances may still be the best one for those specific applications, but particularly for the context of prioritising between interventions in different contexts, it is important to have the ability to consider and compare a wider range.

Scale type

For most of the purposes discussed, such as policy analyses and comparative suffering estimates, it is important to have a cardinal output. That is, that the measure is performed on one of the cardinal measurement scales (interval or ratio), rather than a merely comparative ordinal scale. While there are other applications for which comparative ordinal rankings may be sufficient or even preferred, for the purposes described in this paper, the calculations will require cardinal data. I have argued elsewhere that subjective welfare is measurable on these types of scale (Browning, 2022), but it is important that we choose a measurement method that produces output meaningfully represented on these scales.

The measure should also be bidirectional, capable of representing welfare states in both directions (positive and negative). Some measures are particularly concerned with suffering and do not have room to consider positive welfare experiences, which will skew results. A measure that fails to range across both positive and negative welfare experiences will fail to capture everything we care about. This does not mean that the total possible intensity on either side of the zero point must be the same – it is possible, for instance, that the worst possible states of suffering are worse than the best possible states of pleasure are good and we might want to have something like the +10 to -25 scale used by Norowitz (in Warren 2018) for this reason. All that is required is that the measure can capture experiences on both sides of the neutral line.

Informativeness

It is also preferable for our measures to be informative, in terms of providing information about the particular housing and husbandry conditions that are impacting on the subjective welfare of the animal. This then allows the measurements to be used in guiding action to improve the welfare of these animals. It is only through knowing which conditions are the primary causes of poor (or good) welfare, that decisions can be made regarding what to change.

Feasibility criteria

Finally, there are the considerations of feasibility. We want our measures to be correct and useful, so that they give us accurate results that we can apply where we need them. Both these sets of criteria describe the outputs of the measures, and their fitness for purpose. By contrast, the feasibility criteria refer to the process of measurement, and how easy the measure is to collect and apply across the range of circumstances of interest. These criteria are less important than either of preceding two sets; they would be good to have where possible, but not essential. They can still, however, provide reasons to prefer some measures over others, particularly in the real-world circumstances in which they will be used, with various constraints and limitations.

Ease of use

Ease of use refers to how easy the measure will be to collect and apply. All the measures need to be taken and applied in real-world situations, with limitations on time, money, access to animals etc. This means it is going to be better to have a measure which is easy to collect, preferably a simple procedure that does not require a large amount of time or money. Particularly for large-scale applications requiring measurement of a large number of animals, or for a large range of institutions, time-consuming or complex measurements and calculations may prove intractable. However, in cases of assessing and comparing the typical life quality of animals across different institutions and housing types, it may be possible to instead test a representative sample and extrapolate from there.

Current data availability

One restriction on measures for use now, or in the near future, is current availability of relevant data. Many of the measures I will discuss are quite new, and data is not yet available for many species. In cases where it is important to quickly start making comparisons for immediate action, such as charitable investments or interventions, it may be preferable to choose a measure for which a lot of data has already been collected, rather than one that still holds the requirement for assessors to go into the field and undertake the relevant measurements.

Assessing measures of welfare

I have here described a framework for assessing different measures of welfare with a number of criteria for the measures to meet, taking into account considerations of correctness, usefulness and feasibility. The next stage is then to look at different measures of different types, to assess how well they meet these criteria, as I will do in the sections that follow.[4]

I will not attempt to run a quantitative assessment of the methods against the desiderata. It would be possible to try and score each measure according to how well they meet each of the desiderata and use the resulting tally to choose the 'winner' (see e.g. Charity Entrepreneurship, 2018). However, there is a concern that this sort of method could lead to misleading precision; where meaningful assessment of the items is replaced by imprecise quantification that cannot be checked or validated. Scores would be assigned with a large degree of subjectivity, and the weightings between them would also be highly arbitrary; only with a principled and reliable way of assigning scores and setting weightings would such an approach be appropriate.

Here I have instead used a qualitative approach in considering whether measures meet the criteria. There are no explicit scores given, and no specific weightings applied for the different criteria, though some are given higher priority than others - for example, validity is a necessary condition for a good measure, while data availability is merely preferable. This means there is no definitive rating of the different

[4] See Bridgwater (2021) for a similar approach, using a different set of criteria.

measures, and which is the best for task will depend on contextual factors in the application. The approach instead allows for a discussion of each of their benefits and drawbacks, and of which features the 'ideal' measure should possess. In the following sections I will look at a range of different welfare measures and discuss how they perform in relation to the criteria I have presented above for measuring subjective animal welfare.

The measures are divided into two categories – whole animal measures and combination measures (based on a similar distinction made by Beausoleil and Mellor (2011) between whole animal profiling (WAP) and systematic analytical evaluation (SAE)). Whole-animal measures are a single indicator applied to a single animal, which are taken to represent the entire quality of life as experienced by the animal, at least at the point in time the measure is taken. The degree to which they can represent a longer-term cumulative welfare experience is uncertain. These measures rely on the assumption that an animal is able to internally 'calculate' the balance between different positive and negative affective states and that this produces detectable behavioural and physiological changes representing the output of this process. Justification of this assumption rests primarily on taking the evolutionary role for these affects in guiding trade-offs and decisions for action, and here I will be assuming that such measures can be valid (see Browning 2020 for defence of this claim).

Combination measures are more complex, combining multiple lines of evidence, appropriately weighted to give a single quality of life score. These lines of evidence are all partial measures: indicators that reflect some particular contributor to welfare, such as a specific affect or environmental condition. For example, body condition scoring is often used as an indicator of hunger, or nutritional status. While these partial measures will fail on their own for the task required here, as they are all incomplete, they can be useful when combined for use in some of the frameworks I will describe. These measures also differ from the whole-animal measures as they are often applied at the facility level, to a group of animals, instead of an individual. While in most cases, the frameworks can be used to assess individuals as well as groups, some of the individual indicators may not allow this. Often, if the outputs represent the average welfare across the animals in the facility, they can still be used roughly as individual measures would be. However, where there is a wide range of variation in the individual experiences of animals within the group, this may reduce accuracy. In these cases, we must be careful to pay attention to the context of their use.

Whole-animal measures and combination measurement frameworks thus differ in several of their features. As I will discuss, each of the categories of measure has specific strengths and weaknesses, and as I will argue, are strongest when used together.

Whole-animal measures

The first set of measures I will assess are whole-animal measures. These measures consist of a single indicator, used to represent the total quality of life for the animal. In general, the whole-animal measures are valuable because they can give a single complete score representing the entire subjective welfare state of the animal; and they

are often quick and easy to apply. Their primary drawback is that in most cases they fail to provide information on which conditions in animals' lives are responsible for their good or poor welfare, and thus on their own can't serve as a guide for intervention. In this section I will discuss the most commonly used whole-animal measures – human intuitive estimates, qualitative behavioural assessment (QBA), and cognitive bias – assessing their appropriateness, according to the desiderata.

Human intuitive estimates

As I discussed in the introduction, human intuitive estimates have been a common method for filling in estimates of animal suffering in the calculators used by effective altruism organisations (e.g. Norwood & Lusk, 2011; Tomasik, 2018; Warren 2018). The method involves one or several human observers, who compile information on the life of the animals and conditions they are kept in, and on this basis form a judgement regarding the amount of suffering, or quality of life, of the animal within this system. The time scope of this measure will depend largely on what information is used by the estimators – whether a focus on the current conditions, or incorporating the animals' past – and thus has some flexibility in this regard.

The benefits of this approach – and the reasons for its use so far – are in the usefulness and feasibility. The methods are relatively quick and easy to apply, can be used across a range of species and contexts, and outputs can be placed on whatever type of scale the users choose (though the methods of producing numbers may not justify meaningful cardinal scales). They also provide relevant information on the living conditions of the animals. However, these do not outweigh the problems in meeting the correctness criteria.

The major problem for these methods is the subjective nature of the assessment, based entirely on the intuitive judgements of the observers, which are vulnerable to incomplete information, and anthropomorphic ranking of needs. This is one of the places where the effect of individual values may be strongly present. The subjectivity greatly undermines the correctness criteria for the measure. It is likely to be invalid as what is being measured is not really animal quality of life, but instead something like observer preference for particular kinds of housing situations and types of animal lives. It is also unreliable – as seen in Table 1, there is a large range of variation between different observers and their scores (also seen in comparisons of welfare estimates by Otten et al. 2017; Veasey 2020a), and this means it is highly likely to be inaccurate. The measure may or may not be complete, depending on how well the observer does at incorporating all the aspects of the animals' lives which might impact on welfare.

These methods may be strengthened through use of a Delphi method with a sufficiently diverse panel of experts, to reach consensus on the estimates (Rioja-Lang et al. 2020; Veasey 2020a, b; Whittaker et al. 2021), but would need to be assessed for reliability and validity.

Verdict Perhaps useful as a (very) rough and ready approach for making quick assessments in the absence of any other data - particularly if only trying to rank different systems - but results should be treated with extreme caution. Any detailed

calculations regarding the comparative impact of different interventions are highly unlikely to be accurate.

Qualitative behavioural assessment (QBA)

A more rigorous and more promising version of human intuitive estimates is Qualitative Behavioural Assessment (QBA) (Wemelsfelder et al. 2001). This method holds many of the benefits of the human estimates, without the same drawbacks. In QBA, experienced observers make a judgement about the subjective welfare of animals through direct observation, using the animal's behaviour and body language, and the way it interacts with its environment, as an expression of the total welfare state of the animal. It is an "integrative welfare assessment tool" (Wemelsfelder et al. 2001, p. 209), in which the observer is unconsciously integrating many pieces of information from the behaviour and body language of the animal to form a judgement about its overall mood (Wemelsfelder 1997). It is primarily a short-timescale measure, representing the recent impacts on animal welfare, and thus best used either for assessing the effects of specific immediate changes, or when an animal is viewed when in its typical daily living conditions.

The primary benefits of this method are in feasibility. It allows for a simple and rapid assessment of the wellbeing of an animal, without the need to collect a lot of detailed data. Current data availability is moderate, with the process having been applied to a range of farm animals (Gutmann et al. 2015; Muri et al. 2019; Wemelsfelder et al. 2000; Wickham et al. 2015) and some zoo animals (Delfour et al. 2020; Patel et al. 2019). It gives cardinal outputs that are also bidirectional, identifying animals with both positive and negative overall welfare. Additionally, and importantly unlike the previous methods described, QBA also scores well on correctness criteria. By design, it gives a complete assessment of the entire state of welfare of the animal. It has been validated against other physiological and behavioural welfare indicators (Wemelsfelder 2007), and shows high reliability and accuracy (Fleming et al. 2016).

The primary potential drawback is range of applicability. So far it has mainly been used for large mammals and given its reliance on human estimates of behaviour and body language, it may not be of as much use for species very unlike ourselves or those we are not so familiar with, such as fish and insects. However, it is possible this could be offset through acquiring greater familiarity with different species (Balcombe 2020; Wemelsfelder 2007).

Verdict This method has most of the benefits of the 'human intuitive estimates' approach, without the drawbacks relating to lack of accuracy or validity. However, it potentially has a limited range of use, depending on establishing its validity for a wider range of species – were such range to be validated, it could be a strong feasible method for making quick assessments of overall welfare experience.

Cognitive bias

The final type of whole-animal measure are cognitive bias tests. These measure the overall 'mood' of an animal (representative of its cumulative subjective welfare state) through the effects on cognitive processes (Mendl et al. 2010). The primary test of cognitive bias is judgement bias, which works through identifying the level of 'optimism' or 'pessimism' of an animal, reflective of its mood, or welfare. Individuals who have experienced primarily positive states are more likely to view ambiguous signals optimistically, while individuals who have experienced primarily negative states will be more likely to view ambiguous signals pessimistically. Like QBA, cognitive bias is a more immediate welfare measure and should be applied accordingly.

From tests so far, cognitive bias measures appear to score highly on correctness criteria. They have been validated through analogous work on human cognitive bias, as well as producing the predicted results under experimental manipulation, both environmental and pharmacological (Lagisz et al. 2020; Mendl et al. 2009; Neville et al. 2020), though they have not been specifically tested for reliability. They are a complete measure, taking an 'output' score of the overall mood of the animal, integrating the full range of its welfare experience. They also score well on usefulness criteria. They can give cardinal output scores, based on degree of judgement bias as relative to established maximums and minimums, and these scores are bidirectional, recognising both positive and negative welfare states. They should be applicable across many conditions and species – current work includes mammals (Mendl et al. 2009), birds (Deakin et al. 2016), fish (Laubu et al. 2019) and even honeybees (Bateson et al. 2011).

The primary drawback in judgement bias testing is in feasibility, particularly the advance training required. Not only is this time-consuming, reducing the feasibility of the measure, but may also reduce accuracy as training itself can alter welfare (Roelofs et al. 2016). For this reason, further work into other types of cognitive bias could help develop more suitable tests, which do not require training. These are attention bias, in which animals experiencing negative affect will show increased attention to negative stimuli (Crump et al. 2018), and memory bias, in which animals experiencing negative affect will show greater recall of negative memories (Clegg 2018), as well as measures of anticipatory behaviour, in which an animal will show higher anticipation for reward when in a positive emotional state (Spruijt et al. 2001). Current data availability for these methods is moderate, with some work on a range of farm animals (Deakin et al. 2016; Lee et al. 2018; Scollo et al. 2014) and now some zoo animals (Clegg 2018), but more work is required to produce results from the range of standard housing systems.Verdict This method is probably the most promising of the whole-animal measures, due primarily to the accuracy and range of applicability. Further work is needed to establish the validity and accuracy of the less labour-intensive methods (Table 2).

The above table summarises the discussion of the different types of measures. A tick represents a measure strongly meeting the requirements of the criterion, a cross

Table 2 Assessment of whole-animal measures

	Correctness				Usefulness			Feasibility	
	Valid	Accurate	Complete	Reliable	Range	Scale	Inform.	Ease of use	Data
Human estimate	O	O	O	O	Π	Π	Π	Π	Π
QBA	Π	Π	Π	Π	O	Π	O	Π	-
Cog bias	**Π**	**Π**	**Π**	**-**	**Π**	**Π**	**O**	**O**	**-**

represents failing to meet the requirements, while a dash represents either neither strong failure nor strong success in this regard, or lack of available data to decide. This is intended only as a visual representation of the qualitative assessment above; the measures are not specifically compared on the number of ticks and crosses but on how well they are considered to do across a range of categories, particularly the more important, such as validity. Of the whole-animal measures assessed here, the most promising seems to be cognitive bias, due primarily to its range of applicability; QBA is also a strong contender if it can be shown to be applicable across a wider range of species.

There are some other whole-animal methods that may be promising in the future, such as the markers of biological aging (e.g. telomere length and hippocampal volume), that work on the premise that exposure to stressors will prematurely age an animal, and thus a comparison of the 'biological age' to the actual age will indicate the level of stress the animal has been exposed to, which can be used to infer the quality of life that animal has experienced. These measures reflect the longest timescale, as they represent the total cumulative experience of the animals so far, which makes them potentially the most accurate total welfare indicators but not as well-suited to applications that require a snapshot 'at a time' picture of welfare. The methods have been reviewed elsewhere (Bateson 2016; Bateson and Poirier 2019; Poirier et al. 2019), but primarily still require validation, particularly to ensure that they are tracking subjective animal welfare – both positive and negative - and not just physiological stress. I have not detailed them here, as they are still underdeveloped for the purposes described in this paper, but they may work well if they can be established to meet the criteria I have set out. The framework I have presented is intended for use in just this way – as a tool for ongoing assessment of different methods as they develop.

In general, whole-animal measures are the best way of making an accurate measure of the entire state of subjective welfare for an animal. They will take into account all aspects of welfare and will in general be quicker and easier to apply than combination measures that require multiple lines of evidence. Except for the human intuitive estimates, they are weakest in their inability to provide details about the reasons for the welfare score and thus will do well used in conjunction with the combination measures I will discuss in the next section.

Combination measures

The next set of measures I will assess are combination measures. Combination measures are created using multiple partial indicators, each of which represent a contributor to subjective welfare experience - such as nutrition, health, or behaviour.

These are scored and weighted by their relative contribution to overall experience, to attain an overall quality of life score. They can thus give us detailed information about the impact of different conditions on animal welfare. The major drawback to these models is they risk leaving out some contributors to welfare, leading to incomplete calculations, or that they may have inaccurate weightings between the different components of the model. Like the human intuitive estimates, the timescale scope for these measures will depend on those of the specific indicators used to construct them. Where these are mostly more stable indicators reflecting ongoing living conditions, this will mean the frameworks will provide a good description of the typical daily welfare of the animals, not so sensitive to the impacts of immediate changes.

These types of frameworks are increasingly common, and many are developed specifically for particular species. Here I will assess the most commonly used general combination measurement frameworks - the Five Domains model (Mellor et al. 2020), Welfare Quality protocol (Botreau et al. 2007) and welfare Decision Support System (Bracke et al. 2002; Bracke, Spruijt, et al., 2002). Rather than breaking them down individually, I will discuss the models together, as their relative benefits and drawbacks are best seen as compared and contrasted to one another.

The combination measurement frameworks operate by dividing welfare up into different categories, such as nutrition, housing, health, and behaviour, and identifies different indicators within these to measure different components. The scores for these are then aggregated, first by category, and then overall, to output a single score taken to either represent the welfare of the animal or group, or the quality of the facility as regards the welfare of its animals. These models are highly sensitive to the effects of individual values, as discussed earlier. The selection of relevant categories, as well as the selection of indicators within these categories, will reflect the values and commitments of those building the framework. As I will discuss shortly, this will also be true of the weighting procedures used to aggregate the components into a single score – where this is done by 'expert opinion', it will reflect the choice of experts and their own (often discipline-specific) views about the differing importance of different states.

The frameworks vary regarding their performance on usefulness criteria. Depending on their construction, they are potentially useful across a large range of species and contexts. However, as new sets of indicators need to be developed for each species, the current range of applicability is more limited. While the general domains and associated mental states in the Five Domains will be relevant to most types of animals (as well as captive animals, the model has recently been extended to wild animals: Harvey et al. 2020); Welfare Quality and DSS require new models to be explicitly built for each new species – though Welfare Quality has been used for a range of agricultural animals, including pigs (Czycholl et al. 2016), cattle (de Graaf et al. 2018), and hens (Blatchford et al. 2016) and DSS is currently available for use assessing welfare of breeding sows, (Bracke et al. 2002; Bracke, Spruijt, et al., 2002), chickens (de Mol et al. 2006), cows (Ursinus and Schepers 2009) and salmon (Pettersen et al. 2014; Stien et al. 2013)).

The models also differ on the type of scale used. The Five Domains explicitly uses an ordinal scale (A-E), in order to prevent over-precisification where the data does not support it: "numerical grading was explicitly rejected to avoid facile, non-

reflective averaging of 'scores' as a substitute for considered judgment and to avoid implying, unrealistically, that much greater precision is achievable than is possible with such qualitative assessments" (Mellor 2017, p. 10). However, this limits the contexts of use, as many of the applications described do require a cardinal output. Both Welfare Quality and Decision Support Systems produce cardinal scores, though the use of ordinal scoring on some attributes within the models may undermine the assumption of cardinality for the final output.

The frameworks vary in their feasibility, depending on the methods used for the individual measures within them. While the Five Domains relies primarily on observer ratings of the quality of different aspects of animal housing and care, Welfare Quality and DSS rely more heavily on indicators requiring empirical measurement and there are concerns about the length of time it takes to apply the full Welfare Quality assessment (Andreasen et al. 2013, 2014).

The biggest concerns are with the correctness criteria. These will depend a lot on the specific sets of partial measures used in their construction. While all of the models have explicitly been constructed based on consideration of the subjective experience of animals, most of the individual indicators have not been explicitly validated for their connection to subjective welfare (with the exception of Welfare Quality: Buller et al. 2020; Forkman and Keeling 2009).

Additionally, while these individual measures can all be valid, accurate, and reliable, it is most important that the model as a whole has these features. Only the DSS has been validated, but only against expert opinion, which may be unreliable. Whether a combination measurement framework is valid, or accurate, will depend on two further considerations – ensuring the framework is complete, and that the weightings are accurate. The frameworks will only be complete if all relevant aspects of welfare are covered - where there are missing components, this will mean the measure is incomplete, and also undermine the validity and accuracy. As I will discuss in the next section, confirmation of completeness and validity can perhaps best be achieved through use of whole-animal measures.

The biggest weakness for all frameworks of this type is in setting the weightings for the relative impacts of the different components on subjective welfare experience, and this represents the biggest difference between the different frameworks. The Five Domains framework recognises this problem and does not attempt to compare the relative impact of the different domains, with the end score not intended to be a strong representation of overall quality of life; it is intended rather as a 'focussing' device, to gain a greater understanding of the welfare of an animal, and the conditions impacting it, rather than a measurement tool as such. The aggregation weightings used in Welfare Quality are quite opaque, and seem to be based on expert opinion rather than measured effect on the animals (de Graaf et al. 2018; Sandøe et al. 2019) which means the model as a whole is less likely to be a valid measure of the entirety of welfare experience, as is suggested by the poor correlation of Welfare Quality scores with QBA assessments in cattle (Andreasen et al. 2013).

The strongest framework for facing this aggregation problem is the DSS. In this framework, the attribute weightings are based on information available in the literature. At present, this does not provide much confidence in the weightings used – these were not standardised and what counted as relevant data could vary from weighted

Table 3 Assessment of combination measures

	Correctness			Usefulness				Feasibility	
	Valid	Accurate	Complete	Reliable	Range	Scale	Inform.	Ease of use	Data
Five Domains	Π	O	-	-	Π	O	Π	Π	Π
Welfare quality	Π	O	-	Π	-	Π	Π	-	Π
DSS	**-**	**-**	**Π**	**Π**	**-**	**-**	**Π**	**Π**	**O**

preferences to qualitative comments by scientists in their paper. However, importantly, this is transparent and allows for changes to be easily made as new information is attained (e.g. on range of needs, their link to attributes and the weightings of attributes). The data in the model is directly linked to a table of the referenced data (e.g. comments in scientific papers) to allow for transparency, as well as making it updatable. It is this transparency and explicit capacity to update that gives the DSS its strength as a framework.

Verdict Table 3 summarises the discussion of the combination measures, and how well they meet the proposed criteria. As with the whole-animal measures, this is meant simply as a visual representation of the assessments – the number of ticks and crosses is not a direct reflection of the relative quality of each of the measures. Combination measures are useful as they provide detailed information on the conditions of animal lives, and how they impact subjective welfare, which can be used in providing recommendations for action. Their primary weakness is that they may be incomplete, failing to account for all influences on welfare.

The Decision Support System is the most promising of these frameworks for the purposes described in this paper, with some 'cleaning up' of the inputs – particularly ensuring collection of cardinal rather than ordinal data. This is primarily because it is best able to overcome the potential problems of incompleteness and weighting accuracy through transparency and flexible response to new data. Though Welfare Quality is well-developed for assessing and comparing particular species-specific institutions and housing conditions, it currently has too many subjective judgements built in to be confident about its validity or accuracy for producing a welfare score as is required for the purposes described in this paper. While the Five Domains may be highly effective in making assessments of the welfare conditions present for an animal, without a numerical scoring system it would not be of real use in the contexts discussed in this paper, which require quantitative comparisons.

Conclusions

We have many reasons to want to quantify the welfare levels of animals, including policy decisions, impact assessments, and comparing different interventions. In this paper, I have proposed a range of desirable criteria for a measure of subjective animal welfare, against which I assessed several common welfare measures, to identify which best meet our requirements. In particular, I distinguished between whole-animal and combination measures, which each have different strengths and weaknesses.

In the end, the best option is to use both a combination and a whole-animal measure together, as they have complementary strengths and weaknesses. Whole-animal measures are complete and have higher validity and accuracy, while the combination measures are typically more feasible to apply and give more information about the sources of welfare harms and benefits. Combining them allows us to get a sense of the overall mood/welfare of an animal, while still having sufficient detail about living conditions to allow us to determine where change is required. A similar point is made by Aerts et al. (2006) when arguing for a combined usage of a housing assessment framework, a stockperson evaluation and an animal-based measure to get a complete picture of the animal's welfare. Combined use also allows us to validate the measures against one another to make sure we have not missed anything on either side – for example, the lack of correlation between Welfare Quality scores and QBA assessments in cattle (Andreasen et al. 2013) gives reason to look more closely at each method to determine why they do not agree; and amend or replace the methods accordingly. In particular, lack of agreement may help indicate where combination measures have missed some component of welfare.

As I discussed, one of the biggest weaknesses of the combination measures is the current subjectivity involved in setting weightings for the different components within the model. Without having a way of correctly setting weightings such that they reflect the actual impact different experiences have on welfare from the point of view of the animal, the outputs of the model could be entirely wrong. I suggest that use of whole-animal measures allows us an objective method for determining weightings. We would start by using a whole-animal measure to measure the overall welfare of an animal at one point. We would then make an intervention we were interested in testing the effect of, say by changing food quality or amount of available shelter. Finally, we would measure overall welfare again, to observe the difference in the scores. This difference will help us determine the impact of this condition on overall welfare. Repeating this for many conditions would start to give us their relative weightings. Use of preference tests to see how strongly animals prefer particular conditions over others can also tell us something about their weightings relative to welfare. However, these tests should be used with caution as they will only imperfectly reflect the actual hedonic impact of preferred conditions due to a number of potential confounding factors, most importantly the short-term nature of most preferences (Dawkins 1990; Franks 2019; Fraser and Nicol 2018; Jensen and Pedersen 2008; Kirkden and Pajor 2006).

Having looked at a variety of measures, and assessed them against the desiderata, the current best whole-animal measure of subjective animal welfare is probably cognitive bias, with some more work to ensure its validity and accuracy. The best combination measure will be a DSS framework, as it is the only one of the combination models to have a transparent aggregation system and an objective way of setting weightings. A version of this model, with improved inputs, and with systematic use of whole-animal measures or preference tests to set weightings as described above, will be the best way of creating a complete welfare measure. It also allows for continual updating as we learn more; the primary strength of this type of system. While these are currently the best performing measures, the science of animal welfare is rapidly progressing and there will be continual developments in these and other methods. An

advantage of the framework provided in this paper is that can be used for ongoing assessment of existing and emerging measurement methods, such as those discussed in the end of Sect. Whole-Animal Measures.

Using a measure(s) such as those described above will allow us to quantify subjective animal welfare under different conditions, such as in a dairy farm, an indoor chicken barn, or a wild setting. Accurate measurement of animal welfare is a crucial part of the process of making decisions that include the interests of animals. This paper isn't intended to provide direct guidance on what we should do, but rather to provide better tools for figuring it out. In particular, it requires active engagement with the current science of animal welfare, as well as further scientific and philosophical research to clarify and strengthen our understanding and measurement of welfare. With this work, we get closer to having the information we need to make informed decisions that can reduce suffering and improve animal lives.

Acknowledgements Thanks to the Global Priorities Institute at Oxford University for sponsoring the 2019 ECCP visit at which I developed the first drafts of this manuscript, and the other members, visitors, and conference participants who provided feedback on the ideas presented. Thanks also to the graduate students at ANU's 2019 Kioloa workshop and LSE's Choice group attendees, for helpful questions and comments on presentation of this work. I am particularly grateful to Christian Tarsney, Mark Budolfson, and Jonathan Birch for feedback on earlier written drafts. This manuscript has also benefitted greatly from the detailed and constructive feedback provided by David Fraser and one anonymous reviewer. This research is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, Grant Number 851145.

References

Aerts S, Lips D, Spencer S, Decuypere E, De Tavernier J (2006) A new framework for the assessment of animal welfare: Integrating existing knowledge from a practical ethics perspective. J Agric Environ Ethics 19(1):67–76. https://doi.org/10.1007/s10806-005-4376-y

Alonso WJ, Schuck-Paim C (2021) The comparative measurement of animal welfare: Cumulative time in pain as a universal metric with biological meaning. In *Quantifying Pain in Laying Hens: A Blueprint for the Comparative Analysis of Welfare in Animals* (p. 34). https://tinyurl.com/bookhens

Andreasen SN, Sandøe P, Forkman B (2014) Can animal-based welfare assessment be simplified? A comparison of the Welfare Quality® protocol for dairy cattle and the simpler and less time-consuming protocol developed by the Danish Cattle Federation. Anim Welf 23(1):81–94. https://doi.org/10.7120/09627286.23.1.081

Andreasen SN, Wemelsfelder F, Sandøe P, Forkman B (2013) The correlation of Qualitative Behavior Assessments with Welfare Quality® protocol outcomes in on-farm welfare assessment of dairy cattle. Appl Anim Behav Sci 143(1):9–17. https://doi.org/10.1016/j.applanim.2012.11.013

Balcombe J (2020) Intuition and the invertebrate dogma. Anim Sentience 5(29). https://doi.org/10.51291/2377-7478.1591

Bateson M (2016) Cumulative stress in research animals: Telomere attrition as a biomarker in a welfare context? *BioEssays*, *38*(2), 201–212. https://doi.org/10.1002/bies.201500127
Bateson M, Desire S, Gartside SE, Wright GA (2011) Agitated honeybees exhibit pessimistic cognitive biases. Curr Biol 21(12):1070–1073
Bateson M, Poirier C (2019) Can biomarkers of biological age be used to assess cumulative lifetime experience? Anim Welf 28(1):41–56. https://doi.org/10.7120/09627286.28.1.041
Beausoleil NJ, Mellor DJ (2011) Complementary roles for Systematic Analytical Evaluation and qualitative Whole Animal Profiling in welfare assessment for Three Rs applications. *Proceedings of the 8th World Congress on Alternatives and Animal Use in the Life Sciences, Montreal, Canada*, 21–25
Bekoff M, Pierce J (2017) The Animals' Agenda: Freedom, Compassion, and Coexistence in the Human Age. Beacon Press
Blatchford RA, Fulton RM, Mench JA (2016) The utilization of the Welfare Quality® assessment for determining laying hen condition across three housing systems. Poult Sci 95(1):154–163. https://doi.org/10.3382/ps/pev227
Botreau R, Bonde M, Butterworth A, Perny P, Bracke MBM, Capdeville J, Veissier I (2007) Aggregation of measures to produce an overall assessment of animal welfare. Part 1: A review of existing methods. Animal 1(8):1179–1187
Bracke MBM (2001) Modelling of animal welfare: The development of a decision support system to assess the welfare status of pregnant sows. Wageningen University
Bracke MBM, Metz JHM, Spruijt BM, Schouten WGP (2002) Decision support system for overall welfare assessment in pregnant sows B: Validation by expert opinion. J Anim Sci 80(7):1835–1845
Bracke MBM, Spruijt BM, Metz JHM, Schouten WGP (2002) Decision support system for overall welfare assessment in pregnant sows A: Model structure and weighting procedure. J Anim Sci 80(7):1819–1834
Browning H (2020) *If I Could Talk to the Animals: Measuring Subjective Animal Welfare*. https://openresearch-repository.anu.edu.au/handle/1885/206204
Browning H (2022). The measurability of subjective animal welfare. *Journal of Consciousness Studies* 29(3-4): 150–179. https://doi.org/10.53765/20512201.29.3.150
Budolfson M, Spears D (2020a) Quantifying animal well-being and overcoming the challenge of interspecies comparisons. In: Fischer B (ed) The Routledge Handbook of Animal Ethics. Routledge, pp 92–101
Budolfson M, Spears D (2020b) Public policy, consequentialism, the environment, and nonhuman animals. In: Portmore DW (ed) The Oxford Handbook of Consequentialism. Oxford University Press, pp 591–615. https://doi.org/10.1093/oxfordhb/9780190905323.013.26
Buller H, Blokhuis H, Lokhorst K, Silberberg M, Veissier I (2020) Animal welfare management in a digital world. Animals 10(10):1779. https://doi.org/10.3390/ani10101779
Carlier A, Treich N (2020) Directly valuing animal welfare in (environmental) economics. Int Rev Environ Resource Econ 14(1):113–152. https://doi.org/10.1561/101.00000115
Charity Entrepreneurship (2018), September 17 Is it better to be a wild rat or a factory farmed cow? A systematic method for comparing animal welfare. *Charity Entrepreneurship*. https://www.charityentrepreneurship.com/blog/is-it-better-to-be-a-wild-rat-or-a-factory-farmed-cow-a-systematic-method-for-comparing-animal-welfare
Clegg I (2018) Cognitive bias in zoo animals: An optimistic outlook for welfare assessment. Animals 8(7):104. https://doi.org/10.3390/ani8070104
Cooke S (2021) The Ethics of Touch and the Importance of Nonhuman Relationships in Animal Agriculture. J Agric Environ Ethics 34(2):12. https://doi.org/10.1007/s10806-021-09852-5
Crump A, Arnott G, Bethell E (2018) Affect-driven attention biases as animal welfare indicators: Review and methods. Animals 8(8):136. https://doi.org/10.3390/ani8080136
Czycholl, I., Büttner, K., grosse Beilage, E., & Krieter, J. (2015). Review of the assessment of animal welfare with special emphasis on the "Welfare Quality®;animal welfare assessment protocol for growing pigs" *Archives Animal Breeding*, *58*(2), 237–249. https://doi.org/10.5194/aab-58-237-2015
Czycholl I, Kniese C, Büttner K, Beilage EG, Schrader L, Krieter J (2016) Test-retest reliability of the Welfare Quality® animal welfare assessment protocol for growing pigs. Anim Welf 25(4):447–459
Dawkins MS (1990) From an animal's point of view: Motivation, fitness, and animal welfare. Behav Brain Sci 13(1):1–9

de Graaf S, Ampe B, Buijs S, Andreasen SN, de Boyer des Roches A, van Eerdenburg FJCM, Haskell MJ, Kirchner MK, Mounier L, Radeski M, Winckler C, Bijttebier J, Lauwers L, Verbeke W, Tuyttens FAM (2018) Sensitivity of the integrated Welfare Quality® scores to changing values of individual dairy cattle welfare measures. Anim Welf 27(2):157–166

de Mol RM, Schouten WGP, Evers E, Drost H, Houwers HWJ, Smits AC (2006) A computer model for welfare assessment of poultry production systems for laying hens. NJAS - Wageningen Journal of Life Sciences 54(2):157–168

Deakin A, Browne WJ, Hodge JJL, Paul ES, Mendl M(2016) A screen-peck task for investigating cognitive bias in laying hens.PLOS ONE, 11(7), e0158222

Delfour F, Monreal-Pawlowsky T, Vaicekauskaite R, Pilenga C, Garcia-Parraga D, Rödel HG, García Caro N, Campos P, Mercera B (2020) Dolphin Welfare Assessment under Professional Care: 'Willingness to Participate', an Indicator Significantly Associated with Six Potential 'Alerting Factors'. J Zoological Bot Gardens 1(1):42–60. https://doi.org/10.3390/jzbg1010004

Duncan IJ (2002) Poultry welfare: Science or subjectivity? Br Poult Sci 43(5):643–652

Faria C, Horta O (2019) Welfare Biology. In: Fischer B (ed) The Routledge Handbook of Animal Ethics. Routledge, pp 455–466

Fleming PA, Clarke T, Wickham SL, Stockman CA, Barnes AL, Collins T, Miller DW (2016) The contribution of qualitative behavioural assessment to appraisal of livestock welfare. Anim Prod Sci 56(10):1569–1578. https://doi.org/10.1071/AN15101

Forkman B, Keeling L (2009) Assessment of Animal Welfare Measures for Dairy Cattle, Beef Bulls and Veal Calves. Cardiff University. http://www.welfarequality.net/media/1121/wqr11.pdf

Franks B (2019) What do animals want? Anim Welf 28(1):1–10. https://doi.org/10.7120/09627286.28.1.001

Fraser D (2008) Understanding Animal Welfare: The Science in its Cultural Context. Wiley-Blackwell

Fraser D (2014) Could animal production become a profession? Livest Sci 169:155–162. https://doi.org/10.1016/j.livsci.2014.09.017

Fraser D, Nicol C(2018) Preference and motivation research. In M. C. Appleby, A. S. Olsson, & F. Galindo (Eds.), *Animal Welfare* (3rd Edition, pp. 213–231). CABI

Gutmann AK, Schwed B, Tremetsberger L, Winckler C (2015) Intra-day variation of Qualitative Behaviour Assessment outcomes in dairy cattle. Anim Welf 24(3):319–326

Harvey AM, Beausoleil NJ, Ramp D, Mellor DJ (2020) A ten-stage protocol for assessing the welfare of individual non-captive wild animals: Free-roaming horses (Equus ferus caballus) as an example. Animals 10(1):148

Haynes RP (2008) Animal welfare: Competing conceptions and their ethical implications. Springer

Jensen MB, Pedersen LJ (2008) Using motivation tests to assess ethological needs and preferences. Appl Anim Behav Sci 113(4):340–356. https://doi.org/10.1016/j.applanim.2008.02.001

Kirkden RD, Pajor EA (2006) Using preference, motivation and aversion tests to ask scientific questions about animals' feelings. Appl Anim Behav Sci 100:29–47

Kuruc K, McFadden J(2021) Monetizing the externalities of animal agriculture: Insights from an inclusive welfare function. *Draft Paper*. https://drive.google.com/file/d/1HBXZ3siYxoxCqmGNb4lrzJ0tq7AxCXQJ/view?usp=embed_facebook

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

Lassen J, Sandøe P, Forkman B (2006) Happy pigs are dirty! – Conflicting perspectives on animal welfare. Livest Sci 103(3):221–230

Laubu C, Louâpre P, Dechaume-Moncharmont F-X(2019) Pair-bonding influences affective state in a monogamous fish species. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1904), 20190760

Lee C, Cafe LM, Robinson SL, Doyle RE, Lea JM, Small AH, Colditz IG (2018) Anxiety influences attention bias but not flight speed and crush score in beef cattle. Appl Anim Behav Sci 205:210–215. https://doi.org/10.1016/j.applanim.2017.11.003

Lusk JL, Norwood FB (2011) Animal welfare economics. Appl Economic Perspect Policy 33(4):463–483. https://doi.org/10.1093/aepp/ppr036

McCulloch S, Reiss M (eds) (2017) Animal Welfare Impact Assessment and the Ethics of the Great British Badger Cull [Special Issue]. *Journal of Agricultural and Environmental Ethics*, *30*(4), 465–584

Mellor DJ (2017) Operational details of the Five Domains model and its key applications to the assessment and management of animal welfare. Animals 7(8):60

Mellor DJ, Beausoleil NJ, Littlewood KE, McLean AN, McGreevy PD, Jones B, Wilkins C (2020) The 2020 Five Domains model: Including human–animal interactions in assessments of animal welfare. Animals 10(10):1870. https://doi.org/10.3390/ani10101870

Mendl M, 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(3–4):161–181

Mendl M, Burman OHP, Paul ES(2010) An integrative and functional framework for the study of animal emotion and mood. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1696), 2895–2904

Muri K, Stubsjøen SM, Vasdal G, Moe RO, Granquist EG (2019) Associations between qualitative behaviour assessments and measures of leg health, fear and mortality in Norwegian broiler chicken flocks. Appl Anim Behav Sci 211:47–53

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 Biobehavioral Reviews 108:269–286. https://doi.org/10.1016/j.neubiorev.2019.11.008

Ng Y-K (2016) How welfare biology and commonsense may help to reduce animal suffering. Anim Sentience: Interdisciplinary J Anim Feeling 1(7):1–10

Norwood FB, Lusk JL (2011) Compassion, by the Pound: The Economics of Farm Animal Welfare. Oxford University Press

Otten ND, Rousing T, Forkman B (2017) Influence of professional affiliation on expert's view on welfare measures. Animals 7(12):85

Patel F, Wemelsfelder F, Ward SJ (2019) Using Qualitative Behaviour Assessment to Investigate Human-Animal Relationships in Zoo-Housed Giraffes (Giraffa camelopardalis). Animals 9(6):381. https://doi.org/10.3390/ani9060381

Pettersen JM, Bracke MBM, Midtlyng PJ, Folkedal O, Stien LH, Steffenak H, Kristiansen TS (2014) Salmon welfare index model 2.0: An extended model for overall welfare assessment of caged Atlantic salmon, based on a review of selected welfare indicators and intended for fish health professionals. Reviews in Aquaculture 6(3):162–179. https://doi.org/10.1111/raq.12039

Pierce J (2019) Putting the "Free" Back in Freedom: The failure and future of animal welfare science. In: Dhont K, Hodson G (eds) Why We Love and Exploit Animals. Routledge

Poirier C, Bateson M, Gualtieri F, Armstrong EA, Laws GC, Boswell T, Smulders TV (2019) Validation of hippocampal biomarkers of cumulative affective experience. Neurosci Biobehavioral Reviews 101:113–121. https://doi.org/10.1016/j.neubiorev.2019.03.024

Rioja-Lang FC, Connor M, Bacon HJ, Lawrence AB, Dwyer CM(2020) Prioritization of farm animal welfare issues using expert consensus. *Frontiers in Veterinary Science*, *6*. https://doi.org/10.3389/fvets.2019.00495

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. https://doi.org/10.3389/fnbeh.2016.00119

Sandøe P, Christiansen SB, Appleby MC (2003) Farm animal welfare: The interaction of ethical questions and animal welfare science. Anim Welf 12(4):469–478

Sandøe P, Corr S, Lund T, Forkman B (2019) Aggregating animal welfare indicators: Can it be done in a transparent and ethically robust way? Anim Welf 28(1):67–76

Scherer L, Tomasik B, Rueda O, Pfister S (2018) Framework for integrating animal welfare into life cycle sustainability assessment. Int J Life Cycle Assess 23(7):1476–1490. https://doi.org/10.1007/s11367-017-1420-x

Scollo A, Gottardo F, Contiero B, Edwards SA (2014) Does stocking density modify affective state in pigs as assessed by cognitive bias, behavioural and physiological parameters? Appl Anim Behav Sci 153:26–35. https://doi.org/10.1016/j.applanim.2014.01.006

Soryl AA, Moore AJ, Seddon PJ, King MR (2021) The Case for Welfare Biology. J Agric Environ Ethics 34(2):7. https://doi.org/10.1007/s10806-021-09855-2

Spruijt BM, van den Bos R, Pijlman FTA (2001) A concept of welfare based on reward evaluating mechanisms in the brain: Anticipatory behaviour as an indicator for the state of reward systems. Appl Anim Behav Sci 72(2):145–171

Stien LH, Bracke MBM, Folkedal O, Nilsson J, Oppedal F, Torgersen T, Kittilsen S, Midtlyng PJ, Vindas MA, Øverli Ø, Kristiansen TS (2013) Salmon Welfare Index Model (SWIM 1.0): A semantic model for overall welfare assessment of caged Atlantic salmon: review of the selected welfare indicators and model presentation. Reviews in Aquaculture 5(1):33–57. https://doi.org/10.1111/j.1753-5131.2012.01083.x

Sunstein CR, Hsiung W (2006) Climate change and animals. Univ Pa Law Rev 155:1695–1740

Tomasik B(2015) The importance of wild-animal suffering. *Relations*, *3.2*, 133–152. https://doi.org/10.7358/rela-2015-002-toma

Tomasik B(2018), July 14 *How Much Direct Suffering Is Caused by Various Animal Foods?* Essays on Reducing Suffering. https://reducing-suffering.org/how-much-direct-suffering-is-caused-by-various-animal-foods/

Ursinus W, Schepers F (2009) COWEL: a decision support system to assess welfare of husbandry systems for dairy cattle. Anim Welf 18(4):545–552

Veasey JS (2020a) Assessing the psychological priorities for optimising captive Asian Elephant (Elephas maximus) welfare. Animals 10(1):39. https://doi.org/10.3390/ani10010039

Veasey JS (2020b) Can zoos ever be big enough for large wild animals? A review using an expert panel assessment of the psychological priorities of the Amur Tiger (Panthera tigris altaica) as a model species. Animals 10(9):1536. https://doi.org/10.3390/ani10091536

Veit W, Browning H (2021a) Perspectival pluralism for animal welfare. Eur J Philos Sci 11(9):1–14 https://doi.org/10.1007/s13194-020-00322-9

Veit W, Browning H (2021b) Extending animal welfare science to include wild animals. Anim Sentience 6(31). https://doi.org/10.51291/2377-7478.1675

Warren S(2018), August 22 *Suffering by the Pound: Meat and Animal Product Harm Comparisons*. https://stephenwarrenorg.files.wordpress.com/2018/08/suffering-by-the-pound-meat-and-animal-product-harm-comparisons5.pdf

Wemelsfelder F (1997) The scientific validity of subjective concepts in models of animal welfare. Appl Anim Behav Sci 53(1–2):75–88

Wemelsfelder F (2007) How animals communicate quality of life: The qualitative assessment of behaviour. Anim Welf 16(1):25–31

Wemelsfelder F, Hunter EA, Mendl M, Lawrence AB (2000) The spontaneous qualitative assessment of behavioural expressions in pigs: First explorations of a novel methodology for integrative animal welfare measurement. Appl Anim Behav Sci 67(3):193–215

Wemelsfelder F, Hunter TEA, Mendl M, Lawrence AB (2001) Assessing the 'whole animal': A free choice profiling approach. Anim Behav 62(2):209–220

Whittaker AL, Golder-Dewar B, Triggs JL, Sherwen SL, McLelland DJ(2021) Identification of animal-based welfare indicators in captive reptiles: A Delphi consultation survey. *Animals*, *11*(7), 2010. https://doi.org/10.3390/ani11072010

Wickham SL, Collins T, Barnes AL, Miller DW, Beatty DT, Stockman CA, Blache D, Wemelsfelder F, Fleming PA (2015) Validating the Use of Qualitative Behavioral Assessment as a Measure of the Welfare of Sheep During Transport. J Appl Anim Welfare Sci 18(3):269–286. https://doi.org/10.1080/10888705.2015.1005302