Assessing Measures of Animal Welfare

Heather Browning[1]

[1] School of Philosophy, Australian National University; heather.browning@anu.edu.au

Abstract

When making decisions about action to improve animal lives, it is important that we have accurate estimates of how much animals are suffering under different conditions. The current frameworks for making comparative estimates of suffering all fall along the lines of multiplying numbers of animals used by length of life and amount of suffering experienced. However, the numbers used to quantify suffering are usually generated through unreliable and subjective processes which make them unlikely to be correct. 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 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.

1. Introduction

Around the world, over 50 billion land animals and up to 100 billion fish are raised in intensive ('factory') farms every year, often in very poor conditions. Broiler chickens are kept packed into barns with less than a square foot each, and often suffer deformities and lameness due to breeding for unnaturally fast growth. Salmon - wide-ranging carnivores with complex social lives – are kept within crowded pens on salmon farms, where disease and parasites quickly spread. Female breeding pigs are kept in tiny stalls, without room to turn around, or a soft place to sleep.

With this level of animal suffering in the world, when thinking about how to do the most good, animal welfare is one of the things we should consider. 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 where possible we should act to prevent the former and enable the latter. All else being equal, people would prefer farm animals receive better, rather than worse treatment. But we have limited resources available to us, and so are required to make decisions about where to prioritise our actions or interventions. All our actions come with an opportunity cost, that of some other action we could otherwise have taken that may have provided other benefits. Typically, we want to choose actions in such a way so as we will have the greatest possible impact. If our goal is to reduce suffering, then we want to reduce suffering as much as we can. When considering animal welfare, making decisions like these then requires us to have an accurate understanding of the quality of life of animals living under different conditions. We need to accurately measure animal welfare in order to take relevant and effective action.

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. Here, I take a hedonic view of animal welfare, in which welfare consists in the subjective mental states experienced by an animal - "the quality of its emotional states, including their sign (positive or negative), intensity and duration" (Bracke, 2001, p. 45)[1]. This is a common view within animal welfare science, as well as in writing on animal ethics and welfare within the effective altruism community. 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, and thus are important to measure. Where our goal is not to improve welfare, *whatever that may be*, but instead simply to decrease suffering and/or increase pleasure for animals, then this work will be relevant.

One obvious implication of taking a subjective view of welfare is that we need to believe that animals possess the relevant subjective mental states – i.e. that they are sentient. It is now commonly accepted that at least some animals are sentient, as embodied by the Cambridge Declaration on Consciousness in 2012 – "the weight of evidence indicates that humans are not unique in possessing the neurological substrates that generate consciousness. Non-human animals, including all mammals and birds, and many other creatures, including octopuses, also possess these neurological substrates" (Low et al., 2012, p. 2). A large part of the work in animal sentience is in attempting to identify which animals are sentient and researchers use a range of physiological and behavioural evidence, such as neuroanatomy and response to analgesics (painkillers), to look for signs of sentience. Current research suggests that all vertebrates, cephalopods (octopuses and squid) and arthropods (crustaceans, insects and spiders) are sentient (Ginsburg & Jablonka, 2019). As currently only a small fraction of extant species and even taxonomic groups have been studied to determine sentience, the boundaries

[1] There are a number of important issues in measurement of hedonic welfare that I will not be considering here, such as considerations of the shape of life and integration of positive and negative experiences; these are the subjects of other work.

are still unclear. Whichever animals we find to be sentient will be the targets of our moral concern for their welfare.

Working to identify and implement our best interventions for improving animal lives requires accurate measures of animal welfare. We want to assess the current animal housing systems, in order to identify the best and worst conditions. We also want to compare different possible interventions we might take, to assess their potential effectiveness. A number of writers have been working to this end, to try to produce comparative measures of animal welfare, or – more commonly – animal suffering, with which to compare systems. In animal welfare science, this often takes the form of comparing different farming systems for a particular species, to make recommendations as to the best systems of housing and husbandry (e.g. Botreau, Veissier, & Perny, 2009; Bracke, Spruijt, Metz, & Schouten, 2002). In global priorities and effective altruism research, there has been a growing amount of work - largely unpublished – in creating 'suffering calculators'. These aim to compare the total suffering produced by different types of production systems (e.g. Norwood & Lusk, 2011; Savoie, 2018; Tomasik, 2018; Warren, 2018). 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 some work on wild animals (e.g. Ng, 2016; Tomasik, 2015). In this paper, I am interested in these suffering calculators, and how we can ensure we are inputting the right information to get accurate outputs of comparative suffering with which to guide our decisions.

Suffering calculators have been created to compare the total suffering produced by different sets of conditions – most often different agricultural 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. An example of this type of calculator is shown at Table 1 (taken from Tomasik (2018)). Norwood & Lusk (2011) use an 'Ethical Eating Assessment Tool' which calculates the welfare consequences of a particular action (such as changes in consumption of different animal products) relative to another, through multiplying the number of animals affected by the change by the welfare score of those animals. Scherer et al. (2018) produced a welfare index of 'Animal Life Years Suffered' (ALYS), which is equal to the number of animals used multiplied by measures representing suffering and length of life. 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]).

TABLE 1: Suffering calculator (from Tomasik, 2018)

Animal Product	Average lifespan (days)	Production per lifetime (kg)	Suffering per day of life	Sentience multiplier (0-1)	Number of days of life equivalent to pain of death	Weighted days of suffering (per kg)
Milk	1825	30000	2	0.95	15	0.12
Beef	395	212	1	0.95	30	1.9
Pork	183	65	2.5	1	12	7.5
Turkey	133	9.6	3	0.8	10	36
Chicken	42	1.9	3	0.8	10	66
Cage eggs	501	16	4	0.9	7.5	110
Farmed salmon	639	2.0	1.5	0.5	15	250
Farmed catfish	820	0.39	1.5	0.5	15	1600

The information in these tables is then used to create estimates of the total suffering produced by different systems. For example, we can see in Table 1 that this calculator tells us that 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 (see further discussion in Section 5). These calculators can also factor in effects such as elasticity in purchasing to allow for more fine-grained comparisons between the effects of decisions to switch between different animal products.

These suffering calculators have several components in common. The first is number of animals used. This measure can be taken as a whole measure of simply the total number of animals present within each type of system. Alternatively, and more commonly, it is taken as a production measure; that is, a measure of the production output of each animal used. This is often done in terms of just kg of meat/eggs/milk produced, but also sometimes in slightly more sophisticated ways, such as comparing per gram of protein, or kilojoule count, to allow for nutritional differences between product types. This type of measure helps to ensure equivalence if considering dietary shifts. That is, if a person is planning to reduce pork consumption in favour of chicken, this will result in a different number of animals being used to produce the same volume of food and so we cannot simply compare a single chicken to a single pig in terms

[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/

of quality of life. Actions which reduce consumption of particular products will save a different number of animals, depending on the product. This would be a difficult measure to compare in non-agricultural systems of animal use, such as horse-racing or zoos. In these cases, the simpler 'number of animals used/affected' may be more appropriate.

The second component is lifespan. Length of life is important as it affects the duration of suffering experienced. It seems intuitively compelling that more days of suffering are worse than fewer days of suffering, and more days of pleasure better than fewer. However, this may sometimes produce counterintuitive results, as it will often come out that many animals with low quality of life are better off the shorter their lives are and so this will favour production systems with a high kill rate (though these typically contain a higher number of animals and so will still lose out all things considered due to the high total amount of suffering).

The primary measure of interest for this paper is welfare, or quality of life measure, often recorded as average suffering. It is a measure of the welfare level of the animal; the amount of pleasure or suffering it experiences. This is typically taken as a measure of the quality of the average day experienced by the animal, but can also be modified to take into account the impact of rare or unusual experiences, such as veterinary procedures, handling and slaughter. This can be applied as a relative score, as in the Tomasik (2018) calculator, where beef cattle are assigned a '1' as the lowest level of suffering and others are multiplied from there (so for example, a cage-housed egg-producing chicken is scored as having a life four times as bad as a beef cow). It can also be an absolute score measured on a scale relative to the maximum and minimum possible welfare levels an animal could experience. This is most commonly done on a positive-negative scale such as -10 to 10, where -10 represents the worst possible welfare experience, 0 a neutral experience with a predominance of neither positive nor negative states, and 10 the best possible experience of a happy animal.

Two further components contained in many suffering calculators are sentience multipliers and badness of death measures. Sentience multipliers refer to the relative sentience or capacity for suffering that different species may possess. Badness of death measures try to quantify how bad the loss of life itself is for animals killed prematurely. Both of these are important for comparative total measures of lifetime animal welfare, but too complex to be discussed in detail here, forming the basis for future investigation.

In this paper I will only be concerned with how we fill in the quality of life, or welfare component. 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 2 (adapted from Warren, 2018) shows the estimates given by a few of the more commonly used models, and shows 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). 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.

We want to ensure we have accurate inputs so that we can find those actions which will actually do the most good. In particular, this means we need a good measure of animal suffering, or animal welfare more generally, to enter into these calculators. In this paper, I will begin in Section 2 by looking at the desiderata for a good measure of animal welfare, grouped into the categories of correctness, usefulness and feasibility. I will then go on in Sections 3 and 4 to assess a range of possible methods of measurement according to these desiderata, with recommendations as to which are likely to give us the best results[4]. I will finish in Section 5 by looking at the upshots of these considerations and identifying some useful areas for future work.

TABLE 2: Suffering estimates (adapted from Warren, 2018)

Farmed Animals	**Warren**	**Norwood**	**Shields**	**Norowitz**	**Tomasik**	**Scherer et al.**	**Savoie et al.**
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

[4] A similar project has been started in unpublished work by Savoie (2018). This is a promising start in recognising the need for improved measurement and the value of critically assessing the current possibilities. However, their final framework seems limited, largely due to capturing more than was perhaps intended (i.e. moral value as opposed to welfare), and in the non-transparent weighting procedure of different components.

2. Desiderata for a welfare index

When trying to decide which is the best measure to use in quantifying quality of life for ethical decision-making, we need to have in mind what features this measure must have. The potential criteria can be roughly grouped into three categories – correctness, usefulness and feasibility. Here I will go through a list of proposed desiderata for a welfare index and discuss why each is important for our purposes before moving on in the sections that follow to look at how well different methods of measuring welfare meet these criteria.

2.1 Correctness criteria

Correctness represents the degree to which the measure will give us the right results to use in our calculations, numbers that really do reflect the welfare as experienced by the animals. These are the more crucial criteria, as without the right inputs, our results will be meaningless.

2.1.1 Validity

A measure is valid if it is measuring the intended target, instead of some other state. This is probably the most important of the desiderata, as if a measure is not valid – if it is not actually measuring animal welfare - it does not much matter how well it meets the other criteria. It is thus important that we are very clear about what our target state is – the integrated set of mental states that constitute welfare – to ensure we are measuring this and only this. As mentioned previously, under different conceptions of welfare, there may be other things one thinks comprise welfare (such as some objective states of health or naturalness), but 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. There are a range of things that humans might think are important for or about animals that the metric would then need to capture, despite them not actually relating to welfare as experienced by the animal. This would then give us misleading results, and lead to recommendations of actions which may not actually benefit the animals at all. 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. This requires that a causal relationship hold between the target and the measure – the measure must be either a cause of, or an effect of the target. Correlation alone is insufficient, as many unrelated factors may correlate under particular test conditions (Borsboom, Mellenbergh, & van Heerden, 2004). In some cases, correlation may arise from the target and the measure both being a common cause of another state, but here the measure would be tracking the common cause, not the target and thus would not be a valid measure of the target. In these cases, there is a high possibility of additional effects 'screening off' changes in the target from changes in the measure and so we would not get a reliable correlation.

One might be concerned that we cannot validate any measures of subjective welfare, as the intended target – mental states – is hidden from measurement. Instead, all we have are surrogate or proxy measures, that can only ever be validated against one another. There are two lines of reply to this. First, I think there are good reasons to be confident that with multiple lines of the right kind of evidence, tested against one another, we can infer that the best explanation for observed correlations is changes in welfare states. This is particularly true in cases where we can use manipulations in upstream variables (such as husbandry inputs) to create changes in downstream variables (such as animal-based measurement indicators) to establish causal connections. Second, even if we may not have complete confidence in the validity of any measures of subjective experience, we can still rank the proposed measures for validity, based on how well they have been tested and found to correlate with one another. The problem holds equally for all potential measures. Although we might still deny certainty that any of them are really mapping onto mental states, if any of them are then we can identify which are more likely to be doing so.

2.1.2 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 welfare is high, the measured values are high, and the same for medium, low, neutral etc. This includes sensitivity in detecting relevant changes in welfare. That is, that when there are small changes, increases or decreases in an animal's 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, the changes might be quite small though the total impact still large if affecting a large number of animals. 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.

2.1.3 Holism

We want a measure to be holistic – to provide an account of the entire state of welfare of the animal. What we don't want is a measure that represents only some part of the animal's experience, such that some aspects are left out, or overlooked. For example, some measures may reflect only physical health, while not accounting for psychological aspects of welfare. Holism may refer to representing the entire set of mental states for an animal on a given day (as is the case for most of these calculators). It may also refer to its total welfare experience over a lifetime, requiring the combination of several different measures representing different life states/events and their duration. This could include the 'average day' measure, with additional ones for expected days of particularly high or low welfare (e.g. veterinary procedures, slaughter), weighted for their relative impact and duration.

2.1.4 Reliability

The measure should be reliable, meaning it should give consistent results when repeated. The variation between repeated measures should be low. Repetition in this sense can be of many kinds (Czycholl, Büttner, grosse Beilage, & Krieter, 2015), and ideally our measure should be reliable across all of them. Intra-observer reliability refers to the variability in results for multiple repeated measures taken by the same observer. Inter-observer reliability refers to agreement in measures taken by different observers, of the same target. Test-retest reliability refers to the ability of the test of produce consistent results at different times and under different conditions. High variability in results can be offset to some degree by repeated testing to find common patterns (where there is not bias in one direction) but this will make the measures less feasible due to higher investment of time and resources. Where reliability is low, we would have less confidence that our results in any particular test were accurate, or even that our test is valid. In particular, having observer-independence in our measures will increase our confidence in their reliability (as seen in the lack of reliability for the observer-dependent measures collected for Table 2).

2.2 Usefulness criteria

The correctness criteria described above are the most significant for selecting the right measure, as they ensure we have the right results. However, we also want our measure to do well for the task required. Usefulness criteria describe how well the measures fill the role we require them for – how strong the outputs of these measures are in providing the best data for the job.

2.2.1 Range of applicability

Ideally, we want our measure to be useful across a range of different circumstances of interest. Probably most importantly, we would like to use the same measure for all the species we are investigating. Using different measures for different species risks weakening our comparisons. There is a large range of animal species we want to take into account - from large mammals through to the insects and shrimp now used in farming systems - and our measure should be applicable to all of them. We also want a measure that can be used across a range of types of animal usage, from livestock to wild animals, to increase the scope of our decision-making. A measure that is useful only in a small range of circumstances may still be the best one for those particular applications, but particularly for priority-based decisions, we want to have the ability to consider and compare a wider range.

2.2.2 Cardinality

When thinking about what sort of features the outputs of the model should have, we need to consider for what uses they are intended. Some models have produced numbers intended only for ordinal rankings. For example the Norwood & Lusk (2011) model produces numbers are completely comparative. They represent the relative importance of giving up the different food types, rather than an absolute measure estimating suffering; the value of choices rather than of states. The Tomasik (2018) model gives a similar relative ranking, aiming to show how much of any one product creates as much suffering as another. Where we have a need only for relative rankings, to tell us which are the best or worst states, such ordinal measures are fine so long as they correctly reflect the different levels of suffering.

However, we will often want to make more fine-grained decisions than simply "should we care more about salmon farming than egg production?". We may want answers to questions like, "how much worse off are farmed salmon than dairy cows?" or "are we better off supporting an intervention to improve the lives of laying chickens by 10% or the lives of intensively-produced pigs by 15%?". Here the differences between the levels will also be important. In these cases, the outputs of the welfare index must be cardinal, rather than ordinal, in order to allow us to do the calculations we need to assess different interventions. In practice, this means we are likely to want our measure to produce values according to a ratio scale, with consistent intervals between levels, and a meaningful zero point. Identifying the zero point, what it means and how to measure it, is crucial for measuring welfare and making assessments of experiences, or lives, as positive or negative. It is beyond the scope of this paper to explore this issue, but it is an important area for future work.

Our measure should also be bidirectional; capable of representing both welfare states in both directions (positive and negative). Some measures, particularly those only concerned with suffering, will not consider positive welfare experiences and this can skew results. For instance, Scherer et al. (2018) use a 0-1 scale, where 0 represents best possible welfare and 1 worst possible suffering. They then use this as a suffering multiplier, which means even the best possible state is represented as simply the absence of suffering, not the presence of positive experience. As animals can experience positive welfare, and this is also something we care about, we want our measure to have the capacity to capture it. 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 our measure can capture experiences on both sides of the neutral line.

2.3 Feasibility criteria

We want our measures to be correct and useful, so that they give us accurate results that we can apply where we need them. Feasibility represents how easy the measure is to collect and apply across the range of circumstances we are interested in. 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, give us reasons to prefer some measures over others, particularly in the real-world circumstances in which we will be using them, with their various limitations.

2.3.1 Ease of use

Ease of use refers to how easy our measure will be to collect and apply. All of our measures will 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.

2.3.2 Current data availability

One restriction on measures we want to 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. Where we want to quickly start making comparisons for immediate action, we might prefer a measure for which a lot of data has already been collected, rather than one for which we still need to go into the field and undertake the relevant measurements.

2.4 Assessing measures of welfare

I have listed the desiderata that we want our welfare measure to meet, taking into account considerations of correctness, usefulness and feasibility. There are a number of different types of measures of welfare that we then want to assess, to see how well they meet these criteria. What we don't want to do is to run a quantitative assessment. It would be possible to score each measure according to how well they meet each of the desiderata and use the resulting tally to choose the 'winner'. However, there is a concern that this sort of method could lead to misleading precision; an example of "excessive quantification and systematization (in light of the relatively low amount of explanation and checkability), in an area where evidence is uncertain and in-depth critical judgment is necessary to draw meaningful conclusions" (Cotra, 2017). Scores would be assigned with a large degree of subjectivity, and the weightings between them would also be highly arbitrary. Here I have instead used a qualitative approach in considering whether measures meet the desiderata. There are no explicit scores given, and no specific weightings applied for the different criteria, though some are given higher priority than others. This means there is no definitive rating of the different measures. The approach instead allows for a discussion of each of their benefits and drawbacks, and of which features our 'ideal' measure should possess. In the following sections I will look at a range of different types of welfare measures, and discuss how they perform in relation to the desiderata.

I have divided up the measures into two categories – whole animal measures and combination measures. Whole-animal measures are a single indicator 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. Combination measures are more complex, combining multiple lines of evidence, appropriately weighted to give a single quality of life score. A similar distinction between types of measures is made by Beausoleil and Mellor (2011) who divide types of welfare assessment into whole animal profiling (WAP) and systematic analytical evaluation (SAE). They argue that WAP allows for better integration of evidence into a single welfare score but cannot provide information on specific conditions that might impact welfare. SAE is better in identifying areas of welfare compromise and assessing the potential welfare impacts of interventions to provide specific recommendations on welfare improvements. For this reason, they may be most useful when used in combination, as I will recommend here.

3. 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. Whole-animal measures are in general useful because they can give a single holistic score and are often quite quick and easy to apply. Their primary drawback is in failing to provide information on which conditions in animals' lives are causing them suffering. Here I will discuss the most commonly used measures of this type – human intuitive estimates, qualitative behavioural assessment (QBA), cognitive bias, and neuroimaging – assessing their appropriateness, according to the desiderata.

3.1 Human intuitive estimates

Human intuitive estimates are the most commonly-used measure in providing indices of animal suffering, as currently used for global priorities and effective altruism research (e.g. Norwood & Lusk, 2011; Tomasik, 2018; Warren, 2018). This involves an observer (or, ideally, a set of observers, as used for Savoie (2018)) who observe the conditions of the animals, or read a detailed description of the life and conditions, and on this basis form a judgement as to the amount of suffering, or quality of life of the animal within this system. These deliberations often take the form of something like "how bad is it to be this animal?" or "how much would I like/dislike being this animal?". These assessments are offered with varying degrees of transparency. For example, Warren (2018) provides a detailed reason for each of his scores, including all the factors likely to affect welfare and their duration/intensity. Norwood & Lusk (2011) also spend some time describing the relevant conditions and how these affected their score. By contrast, Tomasik (2018) simply applied a multiplier based on intuitions about animal lives without description. Though this was based on assessment of the descriptive evidence about the conditions, the process of arriving at a decision was not laid out.

The major drawback of this process is the subjectivity. The ratings are based on the intuitive judgements of the observers. Even where the reasons for the judgements are described, the overall decision procedure is entirely subjective; based on the priorities and weightings applied by the observer, which are generally opaque. This is particularly vulnerable to incomplete information, and anthropomorphic ranking of needs. This subjectivity undermines the correctness criteria for the measure. It is likely to be invalid as what is being measured here is not really animal quality of life, but instead something like observer preference for particular kinds of housing situations and types of animal lives. As people are not necessarily very good at seeing life from the point of view of another species (for example, imagining the impact of sensory environments we do not ourselves experience), then this will not often lead to correct results. It is also unreliable – as seen in Table 2, there is a large range of variation between different observers and their scores. This also means the measures are unlikely to be accurate, as the range of measures attributed for each system is so variable. The rough estimation of scores is also likely to give low precision, where small changes in conditions may not be reflected by changes in assigned scores. For example, the difference between a 7.0 and a 7.2 is unlikely to be obvious to most people making the estimates. This low degree of reliability and accuracy was also found by Otten et al. (2017) when comparing the estimates of different experts regarding welfare measures for swine and cattle, with professional affiliation playing a strong role. The measure may or may not be holistic, depending on how well the observer does at incorporating all the aspects of the animals' lives which might impact on welfare, but in most cases, it seems likely that at least something will be missed.

The benefits of this approach are primarily in operationalising – usefulness and feasibility. The measure is cardinal (though we might question just how strongly this is true – given the way numbers are estimated and applied, the differences and ratios between them may not be particularly meaningful). Although it has not always been applied as such, the measure can be bidirectional, allowing for measures of both happiness and suffering. It is applicable across a range of systems and species, providing sufficient information is available and observers feel themselves competent to judge the experiences of those species. It is a quick and easy measure to apply, in cases where sufficient information is available about the housing systems and species of interest. There is sufficient data available about the housing and husbandry conditions of most captive animals to apply this method immediately.

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 is highly unlikely to be accurate. Further work to validate these measures against more objective indicators might give some strength to their use, but given the current variation in estimates, this is unlikely to be successful.

3.2 Qualitative behavioural assessment (QBA)

In Qualitative Behavioural Assessment (QBA) experienced observers make a judgement about the welfare of animals through direct observation. They do this by looking at animal behaviour and body language as an expression of the total welfare state of the animal (Wemelsfelder, Hunter, Mendl, & Lawrence, 2001). This differs from the human estimates described above, in which people make judgements based on observations of the living conditions of the animals alone. While human estimates work on the 'inputs' to welfare, QBA uses the 'outputs'. QBA is carried out by having a set of observers assess the overall body language of an animal against either a set of fixed terms or through generating their own descriptive terms. They score the animal for each of these characteristics (e.g. nervous, alert, curious, excited) by marking a point on a line between the 'minimum' and 'maximum' for these traits. These scores are then converted to numerical scores (0-100, relative to a theoretical minimum and maximum) by the researchers and analysed for general patterns. Rather than looking at a single aspect of an animal, it is a whole-body approach that reflects how an animal is interacting with its environment. 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. The aim is to assess not so much the behaviours themselves, but the 'style' of the behaviour, as representative of the animal's overall mood (Wemelsfelder, 1997).

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 consider the many features of that animal's housing and husbandry that may be relevant. It is a quick and versatile measure that can be used in situations where it might be difficult to collect more detailed data (Fleming et al., 2016). Current data availability is moderate, with the process having been applied to a range of farm animals (Gutmann, Schwed, Tremetsberger, & Winckler, 2015; Muri, Stubsjøen, Vasdal, Moe, & Granquist, 2019; Wemelsfelder, Hunter, Mendl, & Lawrence, 2000; Wickham et al., 2015), though so far often looking for the impacts of specific conditions, such as transport rather than quality of daily life. It gives cardinal outputs, though again we might have some concern as to the nature of these numbers and whether they are robust enough to perform the calculations we require. Further analysis of this method would allow use to judge whether this really is the case. It allows for bidirectionality, to identify animals with both positive and negative overall welfare.

QBA also scores well on correctness criteria. It has been validated against other scientific measures of animal welfare, correlating with other relevant physiological and behavioural indicators (Wemelsfelder, 2007), illustrating that it is not merely subjective judgement but observation of real features of the animal. Although it is not objective, this method has been shown to have high reliability, with similar scores produced by different observers and under different conditions, with minimal training required (Fleming et al., 2016). It appears to be sensitive to subtle differences between animals, or over time (Fleming et al., 2016). It gives a holistic assessment of the entire state of welfare of the animal.

The primary drawbacks are that it is not applicable over very many species. So far it has primarily been used for large mammals (and recently, chickens (Muri et al., 2019)). Given its reliance on human estimates of behaviour and body language, it may not be of much use for species very unlike ourselves or those we are not so familiar with, such as fish and insects (though Wemelsfelder (2007) thinks this is possible, just a matter of acquiring familiarity with and skill in assessing these more phylogenetically distant species).

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 will have a limited range of use, as it may not work for all species and if used alone can't tell us anything about the impact of particular conditions on welfare.

3.3 Cognitive bias

Cognitive bias tests measure the overall 'mood' of an animal. The mood of an animal can be thought of as its cumulative welfare state, a sum or averaging of its positive and negative experiences up to that time (Mendl, Burman, & Paul, 2010). There is good evidence that mood state can affect different cognitive processes (Boissy & Lee, 2014), and these resulting cognitive biases can be used as a test of mood, or welfare.

The primary test of cognitive bias is judgement bias[5], which works by looking to see how 'optimistic' or 'pessimistic' an animal is. In judgement bias tests, animals are trained to expect a reward under one stimulus and a punishment under another. For example, if a light activates

[5] Most papers referring to cognitive bias testing of animals use this to mean judgement bias

in the right-hand corner of the room, they will receive a positive stimulus (e.g. food), and if it activates in the left-hand corner of the room they will receive an aversive stimulus (e.g. a loud noise or strong puff of air to the face). They are then presented with an ambiguous signal, such as a light somewhere in the middle of the room, to see how they react – whether they behave as though they are about to receive the reward (optimistic), or the punishment (pessimistic). Individuals who have experienced primarily positive states - i.e. those who are likely to have lived in an environment providing fitness-enhancing rewards - will be likely to view ambiguous signals optimistically, as potential rewards. Conversely, individuals who have experienced primarily negative states (low reward opportunity environments) will be more likely to view ambiguous signals pessimistically, as potential threats (Mendl et al., 2010). "Mood state may thus act as a heuristic device influencing cognitive processes and facilitating appropriate decision-making behaviour" (Mendl et al., 2010, p. 2900). This leads to a cognitive judgement bias varying with background mood. This has been established in human subjects, and is now the subject of a lot of work on animals, including mammals (Mendl, Burman, Parker, & Paul, 2009), birds (Deakin, Browne, Hodge, Paul, & Mendl, 2016), fish (Laubu, Louâpre, & Dechaume-Moncharmont, 2019) and even honeybees (Bateson, Desire, Gartside, & Wright, 2011) across a range of situations. This work is relatively new, but results so far give positive indication that this is an effective measure of overall welfare state.

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 (Mendl et al., 2009). This also gives us reason to be confident in their accuracy. They are holistic measure, taking an 'output' score of the overall mood of the animal, integrating the full range of its welfare experience. Although they have not been specifically tested for reliability, the apparent success across a range of species and study conditions is promising.

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 has ranged through mammals, birds and fish, and in principle the procedure seems like it could also work for invertebrates.

The primary drawback in judgement bias testing is in feasibility, particularly the advance training required in order to assess an animal. Not only is this time-consuming, reducing the feasibility of the measure, but may also reduce accuracy. Training animals can be enriching, improving their welfare (Melfi, 2013), and so measures taken of animals undergoing this training would not be accurate reflections of the welfare of typical animals under their other husbandry conditions. 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, and memory bias, in which animals experiencing negative affect will show greater recall of negative memories (Clegg, 2018). Measures of anticipatory behaviour, in which an animal will show higher anticipation for reward when in a positive emotional state (Spruijt, van den Bos, & Pijlman, 2001), could be another promising cognitive bias measure, and one that also does not require training. Current data availability is moderate, with some work on a range of farm animals (Deakin et al., 2016; Lee et al., 2018; Scollo, Gottardo, Contiero, & Edwards, 2014), 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 ensure the measures really are valid, accurate and reliable, as well as to standardise against potential maximum, minimum and neutral-level scores.

3.4 Neuroimaging

Neuroimaging is a new procedure, and one that has so far not been well-studied in its application to welfare. This method takes scans of animal brains to look for activation in the regions representing positive or negative mental states. Work identifying the regions of the brain responsible for generating positive and negative experience (Berridge & Kringelbach, 2013) could be used to identify the valence and intensity of emotional reactions to stimuli, and possibly in assessing overall levels of pleasure or displeasure at a specific time. This work is still in the very early stages, mostly in humans, however there have been some promising results where both subjective report of intensity of experience, and behavioural responses, correlate with intensity of brain activity (Coghill, McHaffie, & Yen, 2003). Recently, Poirier et al. (2019) have argued that measures of biomarkers for the hippocampus (a part of the brain responsible for learning, memory and emotional regulation) may give us information about the cumulative affective experience, or overall welfare status, of animals. Hippocampal volume, amount of local grey matter and neurogenesis (development of new neurons) have all found to positively correlate with other measures of welfare, such as subjective self-report in humans, and mood in animals (Poirier et al., 2019).

Currently, work in neuroimaging is too new for us to know whether it may be successful as a welfare measure. The work mentioned above suggests it is likely to be valid, indicated through correlation with other measures. If this is true, it may be the strongest measure we can have, as it is accessing intensity of mental states more directly than any other measure. This would make the measures highly likely to be both accurate and reliable. For now, work has been done primarily on pain, and we would need to establish if results hold across a range of mental states, and if these integrate into overall welfare scores, to give us holism. Neuroimaging measures should score high on usefulness, as they can be applied across a large range of species (though further work would be required to validate the measures for animals with brains quite different than our own, such as fish or invertebrates), and can give cardinal outputs, from intensity of activation along a scale.

The biggest problem in using neuroimaging would be feasibility, as it does require expensive equipment and the restraint of animals. As well as taking time and money, restraint could also reduce accuracy, as we may only be measuring neural response to the stress of the current condition (though the hippocampal biomarkers approach discussed above will not have this problem). Although recent work has used animals trained for voluntary entry into the machines (Berns, 2018), this is not likely to be feasible for a large range of species. There is also no real current data availability, so this method is not useful for current decision-making.

Verdict: Currently too early for any real use, but probably a very promising area for future work.

TABLE 3: Assessment of whole-animal measures

	Correctness				Usefulness		Feasibility	
	Valid	Accurate	Holistic	Reliable	Range	Cardinal	Ease of use	Data
Human estimate	×	×	×	×	✓	✓	✓	✓
QBA	✓	✓	✓	✓	×	✓	✓	-
Cog bias	**✓**	**✓**	**✓**	**-**	**✓**	**✓**	**×**	**-**
Neuroimaging	✓	-	-	-	✓	✓	-	×

The above table summarises the discussion of the different types of measures. A tick represents a measure strongly meeting the requirements of the desideratum, a cross 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, provided further testing confirms it to be a valid and reliable measure. QBA could also be very useful if it can be applied across a wider range of species. Future work on neuroimaging may make this a strong preferred measure.

In general, whole-animal measures are a very useful way of making an accurate measure of the entire state of 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. 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 combination measures.

4. Combination measures

The next set of measures I will assess are combination measures. In creating combination measures, we take multiple indicators, which each represent a component of quality of life, such as nutrition, health and behaviour. We would then take a score for each and weight them according to 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 that we may risk leaving out some aspect of welfare, so that our calculations are incomplete, or may not know enough to produce accurate weightings of the different components. Here I will assess the most commonly used combination measures - the Five Domains model, Welfare Quality® protocol and SOWEL-type models.

4.1 Five Domains

The Five Domains model (Mellor, 2016) measures overall welfare through measurement of each of its components. In this model, the animal is scored for welfare impact over five different domains – four 'physical' domains which reflect the life conditions that affect welfare (nutrition, housing, health and behaviour), and the fifth 'mental state' domain which is what welfare consists in. The four physical domains affect welfare only through their effects on the mental states in the fifth domain. So, for example, the domain of Nutrition may include a component requiring sufficient food quantity, but the ultimate impact of this condition is on the subjective feelings of hunger or satiety, as represented in the fifth domain. The domain of Behaviour may include a component referencing natural foraging behaviour, but the ultimate impact is on subjective experiences of reducing frustration and fulfilling curiosity and satisfaction. This fifth domain counts both positive and negative mental states relating to provisions and lacks in the other four domains. All conditions are therefore represented with their relevant mental states in mind.

For each of the first four domains, the relevant impacts are identified and the animal is assigned a (qualitative) grade A-E which represents the level of compromise in that particular domain (A being no compromise, E being severe compromise), as relating to the negative affects experienced, and their severity, intensity and duration. The fifth domain of mental state then makes some attempt to combine these different impacts in terms of their effects on overall mental state to give a score in this domain. It is in this step that something like an informal integration procedure is used, albeit one based primarily on the knowledge and intuition of the assessor/s. The scores for each domain are brought together and assessed contextually to give a single overall welfare score.

This is a comprehensive welfare measure. It should be valid, as the framework is applied deliberately with animal mental states in mind as the target. There is a degree of subjectivity in the scoring mode, which could compromise reliability, though this has not been tested. This model does not attempt to compare the relative impact of the different domains (e.g. whether health status affects welfare more than nutrition), and so the 'overall' score is not supposed to be a strong representation of overall quality of life, meaning it may not be holistic. It also does not claim to necessarily include all factors that affect welfare, only those for which there are reliable indicators that can be used (Mellor, 2017). This means that we may not have accurate measures, or ones which are sensitive to all changes.

The framework will be highly feasible. It can be applied quite easily as it does not require any specific measurements to be taken; instead observers rate the quality of different aspects of animal housing and care. The information required about the housing and husbandry conditions of the animals is currently available, and so the framework could be quickly applied where needed. It is useful in that it has a high range of applicability – the general domains and associated mental states will be relevant to most types of animals.

The major drawback of the Five Domains framework is that it is an ordinal system only (Beausoleil & Mellor, 2015), where each of the domains is scored qualitatively. The intervals between the grades are not standardised (i.e. the difference between A and B does not necessarily represent the same level of welfare change as between C and D). This was deliberately built into the model 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). This means however that there is no way to combine these scores as there are no weightings given to the different components. Scores can thus only be compared for rank, while not containing any information about the amount of difference between ranks. It is targeted primarily 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. It would not be simple to convert this to a cardinal measure, as the subjective qualitative assessments would not lend themselves to an accurate numerical representation, without giving rise to the problems of the ‘human intuitive estimates’ method discussed earlier. If it were possible to add quantitative information, this could be quite powerful for welfare assessments due to its comprehensive nature.

Verdict: This is a comprehensive model, which maintains focus on the mental states that constitute welfare, but without a numerical scoring system would not be of real use in the context discussed in this paper, which requires quantitative comparisons. Adding a quantitative component to the framework could overcome this limitation but would require substantial reworking.

4.2 Welfare Quality®

The Welfare Quality® framework (Botreau et al., 2007) works on a similar principle as Five Domains, that of using a variety of measures of different aspects of housing and husbandry to create an overall welfare measure. It takes 12 welfare criteria, across four principles – feeding, housing, health and behaviour. Each of the criteria is then assigned several indicators and multiple measures are taken. These are aggregated into the criteria, followed by another aggregation into the welfare principles and finally bringing it all into an overall assessment of the quality of the facility, with regard to the welfare of its animals (Botreau et al., 2009).

This framework does poorly on correctness criteria. The measures are objective and chosen to be repeatable and reliable, so should do well on this front (though see Czycholl et al., 2016 for concerns). Unlike the Five Domains, there is a reduced focus on the affective states of animals and so what we are measuring may not be representative of the subjective welfare states of the animals, meaning this is unlikely to be valid. This also means that some criteria might be missing, or unnecessary, and so also unlikely to be accurate. Missing or incorrect criteria will also mean the measure is not holistic, as there may be parts of welfare experience which are not accounted for. Additionally, the 12 criteria were taken from discussions with consumer-citizens as well as scientists and so do not necessarily represent welfare from the animal point of view. The aggregation weightings are also quite opaque, and seem to be based on expert opinion rather than measured effect on the animals (de Graaf et al., 2018). This means that the model is much less likely to be valid or accurate. This could potentially be overcome through further work to validate the framework against whole-animal measures, but may require substantial changes to the current inputs and weightings.

In terms of usefulness, the framework does well in creating cardinal outputs. However, as new indicators need to be developed for each species it also has limited range of applicability for now, though could be extended to a variety of species in future work. For feasibility, the measures have been selected specifically to be easy to use and apply. There is limited current data availability for many species, though the framework has been used for a range of agricultural animals, including pigs (Czycholl et al., 2016), cattle (de Graaf et al., 2018) and hens (Blatchford, Fulton, & Mench, 2016).

Verdict: This model currently has too many subjective judgements built in to be confident about its validity or accuracy, though further work to validate it could help with this.

4.3 SOWEL-type models

The SOWEL model was developed by Bracke et al. (2002 a; 2002 b) for assessing the welfare of breeding sows. The same framework has been used to create similar models for assessing welfare of chickens (the FOWEL model - De Mol et al., 2006) and cows (the COWEL model - Ursinus & Schepers, 2009). Similarly to the other frameworks described, these models are built by selecting a number of needs (in these models, 12) and scoring various attributes for each need (25 attributes in total for these models). Each attribute is given a discrete score between 0-1 depending on the quality of the measure (e.g. for number of eating places, the categories might be: sufficient, limited, and restricted, receiving scores of 1, ½ and

0). The attributes are then given weightings, based on information available in the literature and a total score calculated as the weighted sum of the attribute scores. These weightings are not just assigned to the whole attribute, but different values within an attribute so that, for example, extreme pain may have a much larger 'pull' on total welfare score than mild pain. This allows for variation in the two types of weightings – weightings between conditions and weightings between different levels of a condition (Norwood & Lusk, 2011). This is a comprehensive procedure and allows for changes to be made as new information is attained (e.g. on range of needs, their link to attributes and the weightings of attributes): "the decision support system is designed to be adaptable, that is, new insights can be incorporated when these become available" (Bracke, Metz, et al., 2002, p. 1835). 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.

These models run into many of the same problems as the Welfare Quality® framework, but the transparency of the aggregation procedures and the capacity to update input data make it much more flexible in the face of new information. Again, the reduced emphasis on mental states may mean that the measures are not valid, tracking other aspects of the animal's life than subjective welfare. Though only currently validated against expert opinion, which may be unreliable, the authors acknowledge that further validation is desirable: "We have only validated the model with expert opinion, which, though not a truly independent measure, is the best available standard at present. Empirical validation would be welcome, but it is presently not known how to measure the overall welfare status directly or indirectly" (Bracke, Metz, et al., 2002, p. 1844). For the same reasons, the framework may have low accuracy. In particular is a problem in the scoring of the attributes – ordinal scores are applied and treated cardinally, without any justification for a linear relationship between the different levels of score (e.g. that the difference between sufficient eating places and limited is the same as between limited and restricted). Given that the categories are arbitrarily chosen, this assumption of cardinality may not often hold. Another issue is in the assignment of weightings – these were not standardised and what counted as relevant data could vary from weighted preferences, to qualitative comments by scientists in their paper. The framework is intended to be holistic, but this will only be true if all relevant criteria have been covered. However, as mentioned, one of the benefits of these models is the transparency, and the ability to update and alter these scores and weightings as new data emerges to improve the framework on all these counts. Although the current inputs into the model are not ideal, the underlying framework of the model is such that this could be easily improved and strengthened.

This model can be highly useful. The data output is cardinal, though not bidirectional, as scores are taken from 0-1, which doesn't allow for tracking of positive and negative welfare states. This would need to change in order for these models to be applied in the range of contexts we are interested in. We must also be careful with cardinal use of the data, as the transformation of ordinal inputs into cardinal outputs is not currently justified. The models can be applicable across a large range of species and situations, providing the relevant frameworks were constructed. It can also be feasible. With the right selection of attribute indicators, measurement could be quite simple, and objective. Although the current availability of data is low – as mentioned above, applied only to sows, cattle and chickens – there is the potential to apply this framework widely.

Verdict: A very promising framework, and the one that seems most likely to work in this setting. Needs further work to strengthen the attribute measures and weightings, such as validating against whole-animal measures and taking relative preferences, but the flexibility of the framework to allow such changes is its greatest strength

TABLE 4: Assessment of combination measures

	Correctness				Usefulness		Feasibility	
	Valid	Accurate	Holistic	Reliable	Range	Cardinal	Ease of use	Data
Five Domains	✓	×	-	-	✓	×	✓	✓
Welfare Quality	×	×	-	✓	-	✓	-	✓
SOWEL	**-**	**-**	✓	✓	**-**	**-**	✓	×

Table 4 summarises the discussion of the combination measures, and how well they meet the proposed desiderata. 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 welfare, which can be used in providing recommendations for action. Their primary weakness is that they may not take into account all influences on welfare. This can be ameliorated to some degree by ensuring we have a full list of conditions in a well-validated model, but still cannot account for internal influences such as individual personality or history. The SOWEL-type models seem to be the best of the set, as they have the ability to provide a full set of conditions with weightings set by data about impact on animals, as opposed to subjective opinion or guesses. Although they are currently not ideal with regard to the data inputs, their strength is in their transparency in terms of the data input and aggregation procedures, and the fact they can easily be modified as new data are discovered.

5. Conclusion

In this paper, I have proposed a range of desiderata for a measure of animal welfare, and then assessed a range of welfare measures against these criteria to find which will best meet our requirements. In the end, what we will want to use is a both a combination and a whole-animal measure together, as they have complementary strengths and weaknesses. Doing so 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. It also allows us to validate the measures against one another to make sure we have not missed anything on either side.

One of the biggest weaknesses of the combination measures is the current subjectivity involved in setting weightings for the different components within the model. Use of whole-animal measures allows us an objective method for determining weightings, to figure out what impact different experiences have on welfare from the point of view of the animal. 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.

Having looked at a variety of measures, and assessed them against the desiderata, I think the current best of the whole-animal measures is cognitive bias, with some more work to ensure its validity and accuracy. The best of the combination measures I think will be something like the SOWEL model, as it is the only one of the combination models to have a transparent aggregation system and an objective way of setting weightings. Something like this model, with improved inputs, and with systematic use of preference tests or whole-animal measures to set weightings, 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.

Using a measure/s such as those described above will allow us to quantify the amount of suffering or quality of life of an animal on a typical day within a particular set of conditions, such as in a dairy farm, an indoor chicken barn, or a wild setting. However, this is not all that is needed to make decisions about our treatment of animals. There are a number of other considerations that need to be taken into account, and further work in these areas will help in setting priorities for action with regards to animal lives. These calculations don't take into account the side-effects of the systems on other animals (e.g. parents or offspring of production animals, or wild animals impacted), which could be important sources of positive or negative value. In most cases, these effects will be highly complex to calculate. The model described here is not intended to stand in as the sole source of data for decision-making, and separate calculations of the secondary costs of production could also be included.

One important area is in making cross-species comparisons, and the use of something like a sentience multiplier to determine their weight in the comparative calculations. This may be based on our credence in the possible sentience of the animal, the moral weight we give it in our deliberations, or empirical data about the relative capacities of different individuals for pleasure and suffering, perhaps based in number of neurons, or complexity of connections. Another is in creating a measure to quantify the badness of premature death for animals. This can be taken simply as the amount of suffering an animal experiences during the capture, handling, transport and slaughter process, but can also include something more to capture the disvalue in loss of life. Again, this is likely to be a highly complex philosophical question, and one that would be fruitful area for further research. As mentioned earlier in the paper, further research in identifying a meaningful neutral or zero point of welfare, would also strengthen the value of the calculations performed in this sort of work.

There is also an important pragmatic question about how we might actually gain access to farms in order to assess conditions and animal welfare. Although most farms, at least within the developed world, must undergo audits, the information gained is not necessarily publicly available. With stricter 'ag-gag' laws limiting the ability of industry outsiders to observe or report on farm conditions, it may be difficult for interested third parties to make assessments. Where conditions are likely to be worst, within the developing world, we are unlikely to have any access at all. Where we believe conditions to be poor, it would be unethical for us to recreate these conditions for animals, simply for the chance to quantify how bad their suffering is. Working together with farmers to gain access to farms and measure and improve welfare is probably the best way forward.

Once we have an accurate quality of life measure in our calculations, we will be much better placed to make decisions about when we should take action on behalf of animals, and of what type. We can compare the value of different interventions before deciding what to do. This work doesn't tell us anything specific about what we should do, but gives us better tools for figuring it out. Some of the types of actions that can be assessed include advocating for changes in production methods to improve welfare and encouraging consumer shifts between or away from animal products to reduce numbers or change types of animals used. In particular, a recommendation arising from this work is that funding would be well-used in improvement of research; to further develop the techniques described for empirically measuring welfare, particularly in areas like cognitive bias testing and neuroimaging.

Accurate measurement of animal welfare is a crucial part of the process of making decisions for animal welfare action. This will require active engagement with the current science of animal welfare. There is also much more scientific and philosophical research to be done 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 and effective decisions to reduce suffering and improve animal lives.

References

Bateson, M., Desire, S., Gartside, S. E., & Wright, G. A. (2011). Agitated honeybees exhibit pessimistic cognitive biases. *Current Biology*, *21*(12), 1070–1073.

Beausoleil, N. J., & Mellor, D. J. (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.

Beausoleil, N. J., & Mellor, D. J. (2015). Advantages and limitations of the Five Domains model for assessing welfare impacts associated with vertebrate pest control. *New Zealand Veterinary Journal*, *63*(1), 37–43.

Berns, G. (2018). *What It's Like to Be a Dog: And Other Adventures in Animal Neuroscience*. Oneworld Publications.

Berridge, K. C., & Kringelbach, M. L. (2013). Neuroscience of affect: Brain mechanisms of pleasure and displeasure. *Current Opinion in Neurobiology*, *23*(3), 294–303.

Blatchford, R. A., Fulton, R. M., & Mench, J. A. (2016). The utilization of the Welfare Quality ® assessment for determining laying hen condition across three housing systems. *Poultry Science*, *95*(1), 154–163. https://doi.org/10.3382/ps/pev227

Boissy, A., & Lee, C. (2014). How assessing relationships between emotions and cognition can improve farm animal welfare. *Revue Scientifique et Technique de l'OIE*, *33*(1), 103–110. https://doi.org/10.20506/rst.33.1.2260

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071.

Botreau, R., Bonde, M., Butterworth, A., Perny, P., Bracke, M. B. M., 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. https://doi.org/10.1017/S1751731107000535

Botreau, R., Veissier, I., & Perny, P. (2009). Overall assessment of animal welfare: Strategy adopted in Welfare Quality®. *Animal Welfare*, *18*(4), 363–370.

Bracke, M. B. M. (2001). *Modelling of animal welfare: The development of a decision support system to assess the welfare status of pregnant sows*. Wageningen University.

Bracke, M. B. M., Metz, J. H. M., Spruijt, B. M., & Schouten, W. G. P. (2002). Decision support system for overall welfare assessment in pregnant sows B: Validation by expert opinion. *Journal of Animal Science*, *80*(7), 1835–1845.

Bracke, M. B. M., Spruijt, B. M., Metz, J. H. M., & Schouten, W. G. P. (2002). Decision support system for overall welfare assessment in pregnant sows A: Model structure and weighting procedure. *Journal of Animal Science*, *80*(7), 1819–1834.

Clegg, I. (2018). Cognitive Bias in Zoo Animals: An Optimistic Outlook for Welfare Assessment. *Animals*, *8*(7), 104. https://doi.org/10.3390/ani8070104

Coghill, R. C., McHaffie, J. G., & Yen, Y.-F. (2003). Neural correlates of interindividual differences in the subjective experience of pain. *Proceedings of the National Academy of Sciences*, *100*(14), 8538–8542.

Cotra, A. (2017, September 15). How Will Hen Welfare Be Impacted by the Transition to Cage-Free Housing? Retrieved June 12, 2019, from Open Philanthropy Project website: https://www.openphilanthropy.org/focus/us-policy/farm-animal-welfare/how-will-hen-welfare-be-impacted-transition-cage-free-housing

Czycholl, I., Büttner, K., grosse Beilage, E., & Krieter, J. (2015). Review of the assessment of animal welfare with special emphasis on the &quot;Welfare Quality&lt;sup&gt;®&lt;/sup&gt; animal welfare assessment protocol for growing pigs&quot; *Archives Animal Breeding*, *58*(2), 237–249. https://doi.org/10.5194/aab-58-237-2015

Czycholl, I., Kniese, C., Büttner, K., Beilage, E. G., Schrader, L., & Krieter, J. (2016, November). Test-retest reliability of the Welfare Quality® animal welfare assessment protocol for growing pigs [Text]. Retrieved May 23, 2017, from http://www.ingentaconnect.com/content/ufaw/aw/2016/00000025/00000004/art00006

de Graaf, S., Ampe, B., Buijs, S., Andreasen, S., Roches, A. D. B. D., Eerdenburg, F. van, … Tuyttens, F. (2018). Sensitivity of the integrated Welfare Quality® scores to changing values of individual dairy cattle welfare measures. *Animal Welfare*, *27*(2), 157–166.

De Mol, R. M., Schouten, W. G. P., Evers, E., Drost, H., Houwers, H. W. J., & Smits, A. C. (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, W. J., Hodge, J. J. L., Paul, E. S., & Mendl, M. (2016). A Screen-Peck Task for Investigating Cognitive Bias in Laying Hens. *PLOS ONE*, *11*(7), e0158222. https://doi.org/10.1371/journal.pone.0158222

Fleming, P. A., Clarke, T., Wickham, S. L., Stockman, C. A., Barnes, A. L., Collins, T., & Miller, D. W. (2016). The contribution of qualitative behavioural assessment to appraisal of livestock welfare. *Animal Production Science*, *56*(10), 1569–1578.

Ginsburg, S., & Jablonka, E. (2019). *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness*. MIT Press.

Gutmann, A. K., Schwed, B., Tremetsberger, L., & Winckler, C. (2015). Intra-day variation of Qualitative Behaviour Assessment outcomes in dairy cattle. *Animal Welfare*, *24*(3), 319–326.

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, L. M., Robinson, S. L., Doyle, R. E., Lea, J. M., Small, A. H., & Colditz, I. G. (2018). Anxiety influences attention bias but not flight speed and crush score in beef cattle. *Applied Animal Behaviour Science*, *205*, 210–215. https://doi.org/10.1016/j.applanim.2017.11.003

Low, P., Panksepp, J., Reiss, D., Edelman, D., van Swinderen, B., & Koch, C. (2012, July). *The Cambridge declaration on consciousness*. Presented at the Francis Crick Memorial Conference, Cambridge, England.

Melfi, V. (2013). Is training zoo animals enriching? *Applied Animal Behaviour Science*, *147*(3–4), 299–305. https://doi.org/10.1016/j.applanim.2013.04.011

Mellor, D. J. (2016). Updating animal welfare thinking: Moving beyond the “Five Freedoms” towards “A Life Worth Living.” *Animals*, *6*(3), 21.

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

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. *Applied Animal Behaviour Science*, *118*(3–4), 161–181.

Mendl, M., Burman, O. H. P., & Paul, E. S. (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, S. M., Vasdal, G., Moe, R. O., & Granquist, E. G. (2019). Associations between qualitative behaviour assessments and measures of leg health, fear and mortality in Norwegian broiler chicken flocks. *Applied Animal Behaviour Science*, *211*, 47–53.

Ng, Y.-K. (2016). How welfare biology and commonsense may help to reduce animal suffering. *Animal Sentience: An Interdisciplinary Journal on Animal Feeling*, *1*(7), 1–10.

Norwood, F. B., & Lusk, J. L. (2011). *Compassion, by the Pound: The Economics of Farm Animal Welfare*. Oxford: Oxford University Press.

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

Poirier, C., Bateson, M., Gualtieri, F., Armstrong, E. A., Laws, G. C., Boswell, T., & Smulders, T. V. (2019). Validation of hippocampal biomarkers of cumulative affective experience. *Neuroscience & Biobehavioral Reviews*, *101*, 113–121.

Savoie, J. (2018, September 18). Is it better to be a wild rat or a factory farmed cow? A systematic method for comparing animal welfare. Retrieved May 28, 2019, from Effective Altruism Forum website:

https://forum.effectivealtruism.org/posts/cimFBQbpjntoBAKCq/is-it-better-to-be-a-wild-rat-or-a-factory-farmed-cow-a

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

Scollo, A., Gottardo, F., Contiero, B., & Edwards, S. A. (2014). Does stocking density modify affective state in pigs as assessed by cognitive bias, behavioural and physiological parameters? *Applied Animal Behaviour Science*, *153*, 26–35. https://doi.org/10.1016/j.applanim.2014.01.006

Spruijt, B. M., van den Bos, R., & Pijlman, F. T. A. (2001). A concept of welfare based on reward evaluating mechanisms in the brain: Anticipatory behaviour as an indicator for the state of reward systems. *Applied Animal Behaviour Science*, *72*(2), 145–171.

Tomasik, B. (2015). The Importance of Wild-Animal Suffering Wild Animal Suffering and Intervention in Nature: Studies and Research Contributions. *Relations: Beyond Anthropocentrism*, (2), 133–152.

Tomasik, B. (2018, July 14). How Much Direct Suffering Is Caused by Various Animal Foods? Retrieved May 21, 2019, from Essays on Reducing Suffering website: 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. *Animal Welfare*, *18*(4), 545–552.

Warren, S. (2018, August 22). Suffering by the Pound: Meat and Animal Product Harm Comparisons.

Wemelsfelder, F. (1997). The scientific validity of subjective concepts in models of animal welfare. *Applied Animal Behaviour Science*, *53*(1–2), 75–88.

Wemelsfelder, F. (2007). How animals communicate quality of life: The qualitative assessment of behaviour. *ANIMAL WELFARE-POTTERS BAR THEN WHEATHAMPSTEAD-*, *16*, 25.

Wemelsfelder, F., Hunter, E. A., Mendl, M., & Lawrence, A. B. (2000). The spontaneous qualitative assessment of behavioural expressions in pigs: First explorations of a novel methodology for integrative animal welfare measurement. *Applied Animal Behaviour Science*, *67*(3), 193–215.

Wemelsfelder, F., Hunter, T. E. A., Mendl, M., & Lawrence, A. B. (2001). Assessing the 'whole animal': A free choice profiling approach. *Animal Behaviour*, *62*(2), 209–220.

Wickham, S. L., Collins, T., Barnes, A. L., Miller, D. W., Beatty, D. T., Stockman, C. A., … Fleming, P. A. (2015). Validating the Use of Qualitative Behavioral Assessment as a Measure of the Welfare of Sheep During Transport. *Journal of Applied Animal Welfare Science*, *18*(3), 269–286. https://doi.org/10.1080/10888705.2015.1005302