ANIMAL ASK

Measuring Animal Welfare

Philosophical foundations, practical indicators and overall assessments

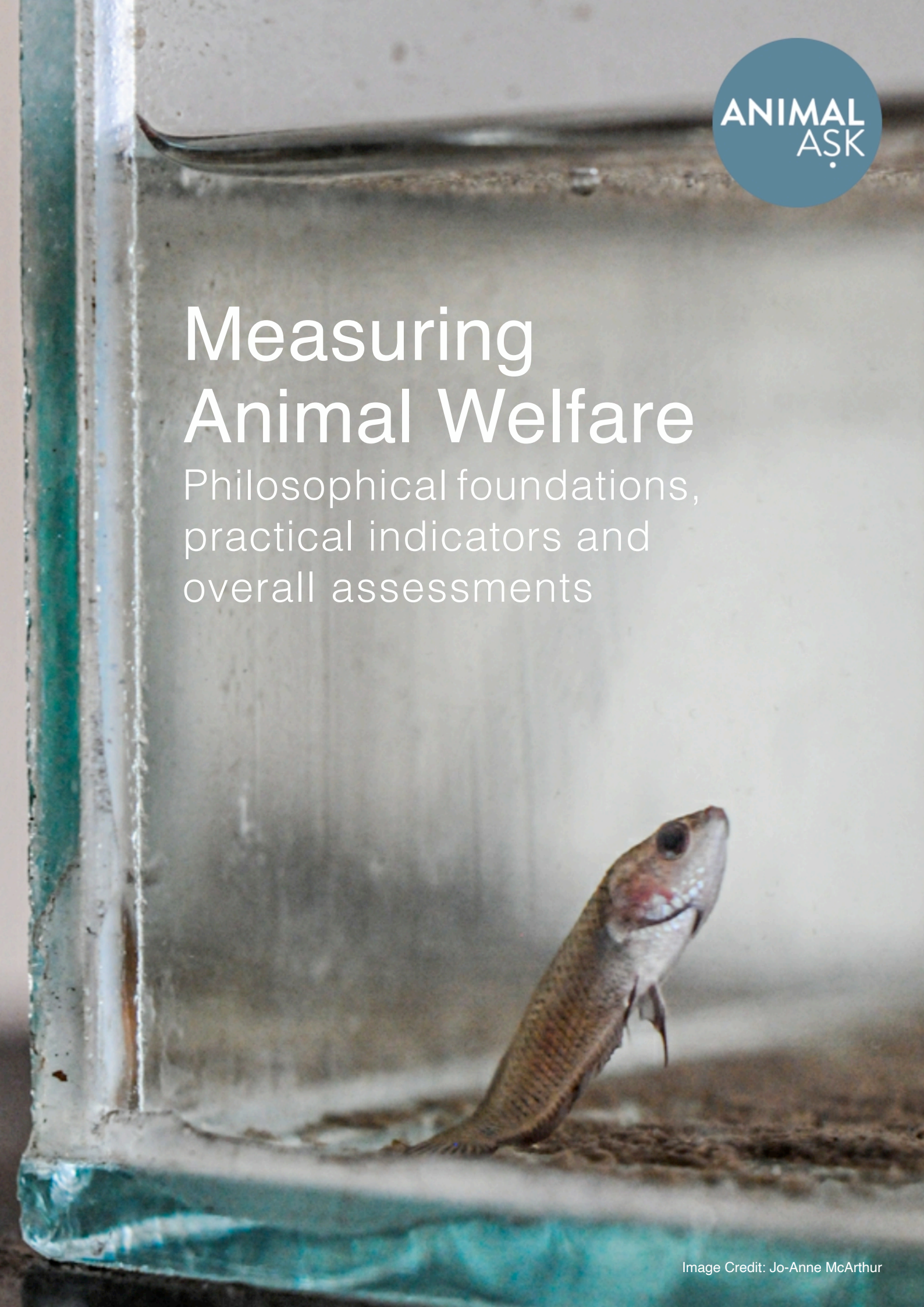


Image Credit: Jo-Anne McArthur

EXECUTIVE SUMMARY

In this report, I examine animal welfare through three stages of analysis. The first is a philosophical discussion surrounding the definition of what animal welfare is and how this definition could differ qualitatively between different animals. This provides the foundation for further research as the validity of practical measures can only be tested with a concrete understanding of the underlying construct. Overall, we accept the definition presented by Bracke and Hopster (2006) that 'Animal welfare is the quality of life as perceived by the animal itself'* but recognise that this still leaves many open questions. In our analysis, we place the most weight on the hedonic and then desire theories of well-being; these value positive and negative experiences or preferences respectively. The relative weight each reader may place on different theories of well-being in non-human animals may vary and should be considered in subsequent sections.

Once the concept of animal welfare has been more clearly defined, I then investigate the strength of different welfare indicators used in the scientific literature. These are the on the ground, testable attributes that one can use to assess the welfare of an individual. In non-human animals, these can be categorised into four clusters: preference tests, physiological indicators, physical health, and behavioural indicators. Each indicator within each category has its own strengths and weaknesses. However, preference tests are the strongest indicators on the whole as they more directly reflect the mental state of the animal. Although they are the easiest to measure, physiological indicators are the weakest measures as they vary depending on numerous other factors or are associated with both positive and negative valence experiences. Therefore, these should be taken in the context of many other indicators to create a holistic picture of the welfare of the individual.

Finally, the report concludes with an examination of current attempts to synthesise these indicators into an overall evaluation of animal welfare. These systems attempt to gather information on various indicators to provide an overall assessment of welfare. These include both the animal-based indicators examined in this report and environmental conditions. Every measure examined has numerous flaws so the results of each system should always be considered in light of its limitations. I argue that rather than relying on any given assessment the best solution is to use a combination of methods that rely on different techniques. The ideal system would use a combination of qualitative measures, expert opinion based measures, an index of animal-based measures, and standalone measures such as preference testing or qualitative behavioural assessment. This combination would provide a variety of qualitative and quantitative perspectives using information for a wide variety of indicators to guide decision making. In practice, where time constraints limit the extensiveness of our research a more limited combination may have to be used.

*It is worth noting that although this is what we perceive to be valuable for an animal's well-being this does not constitute all we find morally valuable for prioritising asks. Other considerations include the probability of sentience (Schukraft, 2020). We are aware that given our moral uncertainty (Open Philanthropy, 2018) one ought to assign some value to other aspects of an animal's life such as avoiding any violation of their rights. Although these are all important considerations in the wider debate about how we should treat animals, this is outside of the scope of this report.

CONTENTS

Measuring Animal Welfare: Philosophical foundations, practical indicators and overall assessments.

Written by George Bridgwater,
Research Lead at Animal Ask, 2021.

Email: info@animalask.org Web: animalask.org

WHAT IS ANIMAL WELFARE?

Image Credit: Luke Stackpoole

The aim of our research is to support decision-makers whose work impacts the lives of animals, so it is important to have a reliable way to measure their welfare. The perfect measure would allow us to evaluate the impact of different environmental factors or life events on the cumulative welfare of an animal's life. This would provide a clear picture of what is most important to the quality of life of the animal and the overall balance between positive and negative welfare events in their life. Unfortunately, this perfect measure does not exist. Even in humans where we can collect self-reports of their global well-being, there is still disagreement about which method is best. We aim to build on the philosophical underpinnings of theories of well-being and existing empirical research on animal welfare to find the most reliable and rigorous measures possible.

Animal welfare is defined differently by many different stakeholders involved in animal agriculture. When discussing animal welfare there are three components people generally refer to;

- biological functioning and health
- the ability of the animal to live a reasonably natural life
- the affective states of the animals (Fraser, 2008)

Each component is weighted differently depending on the reviewer's values and although all of these components seem like reasonable proxies for welfare, different weightings can lead to drastically different conclusions. An example of this provided by Fraser et al. (1997) is of two dog owners, one who places a higher value on health and the other on the dog's ability to live a natural life. The first owner's dog receives regular veterinary care, two meals a day of low-fat dog food and is walked on a leash. The second owner's dog (the owner's third - the first two had been killed by cars) has burns on their coat, is fed generously but sporadically and never wears a collar. The owners place different weightings on specific values and each owner believes that their dog has higher welfare. The underlying true value of the welfare of the two dogs is distorted by what each owner values. We see these differing values in animal agriculture where producers may define high welfare as high productivity, consumers as the ability for the animals to live naturally, and animal advocacy organisations may define high welfare by positive affective states.

At Animal Ask, we focus on what the animals themselves would value and thus accept the definition of welfare proposed by Bracke and Hopster (2006) that: 'Animal welfare is the quality of life as perceived by the animal itself.' However, this leaves open the question of what components an animal might value as part of their quality of life. Here we can draw on the existing discussion on theories of wellbeing. These are mostly focused on humans but offer some key insight into ways in which we can consider animal welfare.

There are three standard theories of well-being, which may offer some illumination; hedonism, desire theories and objective list theories (Crisp, 2017). Hedonism views well-being as the balance of positive and negative experiences. This school of thought would focus specifically on the

maximisation of positive affect for an individual and minimisation or elimination of negative affect in order to achieve optimal well-being. Desire theories view well-being as the extent to which your desires and preferences are met. Finally, objective list theories of well-being consist of a list of generalised values, which might include 'items' such as knowledge, friendship or pleasure, and the extent to which an individual has these goods. Objective goods are said to benefit an individual regardless of whether or not they desire it, or even have positive reactions to it.

HEDONISM

Even though this element of well-being is traditionally focused on in the study of animal welfare there are some notable objections to pure hedonism. The most common objection is the experience machine (Crisp, 2017) which asks us to imagine the option to plug into a virtual reality that would simulate all the best possible experience for the rest of your life. Once in the machine, your experience would be otherwise indistinguishable from the real world and you would not know of ever being put into the experience machine. Many philosophers share the intuition that it would be a mistake to plug in as we value more than just the valence of our experience. Such that an illusion of our preferences or objective lists being satisfied doesn't constitute real well-being. An additional example of this in the animal context is of a cow being upset by the theft of her calf (StJules's, no date). In a purely Hedonism view drugging her into forgetting or being happy with her calves absence is acceptable. However one could argue that this is false well-being as even if we make the cow feel better we are still thwarting her preferences.

DESIRE THEORIES

Heather Browning provides useful criticism of the desire theory of well-being, specifically in the animal context (Browning, 2019a). She argues that preferences, at least in animals, ultimately simplify down to preferences between hedonic states. This is commonly argued when it comes to theories of well-being in humans but she outlines additional reasons to believe it would be the case in non-human animals. The case for desire theory in humans seems to rest on our ability to have higher-order preferences to refrain from indulging a hedonic desire for a preference for some other good. However, Browning argues that in other animals 'without higher-order preferences, it is difficult to imagine exactly what preferences would be if not just the positive association with one state of affairs and the negative association with its frustration'.

The physical states that correspond to desire theories and hedonism seem to be the so-called 'wanting' and 'liking' systems in the brain (Berridge, 2009), mediated through dopamine and opioids respectively. Although these systems track each other with 'wanting' systems generally correlating with 'liking' systems, this is not always the case. Humans can evaluate their overall physical state and determine which states they prefer to maximise in each case. An addict can know that their strong desire to continue using is detrimental to their welfare or a monk that their desire for virtue is more important than pleasure. Even if one can still argue that this is still a preference

for future hedonic pleasure or that satisfaction of being virtuous.

If we look at the question of meta-preferences through this lens then the question becomes which other animals are capable of this meta-cognition and whether the lack of this ability implies that these individuals can't have meaningful philosophical preferences beyond maximising affect. The former is more of an empirical claim as to whether each individual is cognitively sophisticated enough to make these higher-order evaluations. As we are unable to collect self-reports to fully understand most animals' decision making the best evidence we can gather is of their revealed preferences which we can then use to infer their actual preferences. However, the conclusions we draw from these observations is susceptible to different interpretations. For example, if we observe an individual in a zoo repeatedly attempting to escape different observers could interpret this as either an instinctual decision that will lead to poorer welfare, a reflection of what maximises their hedonic well being or a genuine higher-order preference for freedom. Thus any given observation of non-human animals seemingly valuing non-hedonic goods is difficult to interpret as conclusive evidence for the ability to have higher-order preferences.

Even if this question was resolved and we were able to establish certain individuals lacked this ability there is an additional question as to whether the lack of this ability implies that these individuals can't have meaningful philosophical preferences. To illustrate this problem, Imagine the case of a libertarian who desires to be free even if, by some measures, this would reduce their well-being. What characteristic would we have to remove to make this preference insubstantial? It seems possible to remove the ability to do an overall evaluation of their life or meta-cognition and to still have philosophically meaningful preferences. In this case, they might not be able to report that being unfree reduces their well-being, but as they would if they could, the example suggests that being unfree still reduces their welfare.

It is also possible that even if there is a distinct preference system, these preferences are not considered morally valuable or are valuable but don't constitute welfare. An example of this is an addict's preference to continue using a drug that causes them significant pain. Most people would share the intuition that we should ignore such a preference and help them overcome their addiction. However, one could use the same line of reasoning to show that the libertarian's preference for freedom over comfort is also misguided.

OBJECTIVE LIST THEORIES

The objective list theory faces the strongest objection of any of the three theories - the claim that well-being must be experienced. Christopher Rice (2013) defends the theory against this objection through an examination of the 'value of knowledge'– one objective item that he believes contributes to well-being. One example of this in practice is when a doctor tells their patients the truth about their medical conditions even when their patients do not want or need to know their prognosis. He argues that many people intuitively judge that others would benefit by knowing important truths about their lives and are worse off by remaining ignorant. It is

implied that this intuition is evidence that we include things beyond hedonic or desire theories in our understanding of well-being.

However, there are other equally plausible explanations for this intuition. This could either be explained through the desire theory of well-being, in that we tend to assume that the individual ultimately desires to know the truth regardless of its effect on their affect because we tend to assume that other individuals secretly share our preferences, or because value truth, freedom or other objective goods in addition to our well-being but that they don't ultimately make up our well-being.

Although this does not completely negate this intuition it does show that there are alternative explanations for it and so we can't rely on it as strong evidence for objective list theories of well-being.

CONCLUSION

Arguments for each theory of well-being rest on intuitions and thoughts experiments that individuals tend to widely disagree on. So any examination of indicators and evaluation methods of animal welfare should be transparent in the underlying philosophical assumptions of the author. As such, I place the most weight on hedonism and desire theories of well-being and believe that these are what ultimately constitute well-being or animal welfare. If one disagrees with this position either by placing more or sole weight on either component or by including objective list theories then this should be considered in

INDICATORS OF ANIMAL WELFARE

Image Credit: Artem Beliaikin

There has been previous debate as to whether subjective theories of well-being, such as desire and hedonism theories, are a scientific concern. Some argue that as we are unable to directly measure subjective experiences, they fall outside of the realm of scientific enquiry (Fraser et al. (1997). However, even if we are unable to use self-reports of the individual's subjective experience, as we can in humans, there are other measurements we can use to gauge the emotions of others. These measurements are what is referred to as welfare indicators in this report and include real-world measurements that can be made to indicate the welfare of an animal. For example, the subjective experience of fear can be seen through the expression of freezing or fleeing behaviour, alterations in physiology such as changes in heart rate, blood pressure, and circulating glucocorticoids (Mendl et al., 2009). The health of an animal can also be used as a rough proxy for their subjective experience where poor health is often associated with pain and other indicators of poor welfare. As we place some credence in hedonistic and desire theories of well-being and welfare we focus our examination of existing indicators through this lens.

There are two main methods for assessing the strength of psychological measurements used by social scientists: reliability and validity. Reliability is the extent to which measurements are repeatable on different occasions or using different instruments (Drost, 2011). Validity measures the extent to which the measurement measures the variable it purports to measure.

Reliability is used to assess if the measure is consistently measuring the same underlying construct. It allows social scientists to assess the amount of random error for a measurement technique. If random errors are too high, the variance becomes too large to make inferences from small samples. Thus, reliability is necessary for indicators to be useful in any practical sense regardless of their validity.

There are three ways to test reliability: inter-rater reliability, internal consistency and test-retest reliability. Inter-rater reliability measures the level of agreement from multiple experts who provide a rating from the same data or observation. This applies to many health measures of welfare that require experts to rate cleanliness or feather damage (Decina *et al.*, 2019). Internal consistency measures how well multiple items in one measure that are trying to measure the same thing correlate, this is usually evaluated using Cronbach's Alpha.

Test-retest measures are the ‘temporal stability of a test from one measurement session to another’(Drost, 2011). This will be difficult to use for many indicators as the property being measured may vary over short periods.

If an indicator is reliable enough to be operationalised then our next concern is its validity. The overall assessment of an item's validity is known as construct validity which is an evaluation of whether a measurement tool accurately represents the thing we are interested in measuring (Cronbach and Meehl, 1955).

Construct validity is assessed through content, criterion and face validity. Content validity measures the extent that the measure captures the full underlying construct. For example, measures of psychological stress only capture negative experiences rather than the whole range of positive and negative experiences. Criterion (or statistical conclusion) validity examines how well the construct correlates with things we suspect it should be correlated with. Face validity is a subjective judgement about how well the measurement maps onto the property.

In addition to the reliability and validity of the measure, it also needs to be comparable across individuals. This is referred to as Interpersonal comparison of utility (ICU) and is the comparison of the utility or welfare level of two or more individuals. A good indicator of welfare should allow us to reliably compare welfare between individuals. This is important for any indicator or metric to be an effective guide for policy beyond Pareto improvements in welfare, where no individual is made worse off by the change (Kaminitz, 2018; Harrod, 1938). Thus the accuracy of a measure when making ICUs is one of the most important factors to take into consideration.

The issue that arises with ICUs is that utility depends on both an individual's identity and their basket of goods (in this context their environment and experiences). Even on cardinal scales, the upper and lower bounds may vary across individuals. Attempts to transform ICU’s into interpersonal comparisons of utility differences (ICUDs) does not entirely solve this problem, as they rely on equal or similar sensitivity which could also be false. A metric or indicator that makes perfect ICUs should be able to account for both identity and goods removing the problem of utility range.

One proposed solution to the problem of identity is to assume, like most moral philosophers, that ‘deep down we are all alike’(Hammond, 1991). However, our characteristics change from experiences and thus individuals will value baskets of goods differently. This means any environmental-based measure will be slightly inaccurate on an individual by individual basis. For animal-based measures, this will vary depending on the type of measure for some measures we may be more confident in the assumption that shared biology is sufficient to assume indicators are comparable across individuals. Comparing indicators of welfare across different species becomes particularly problematic as we are much less able to make Hammonds assumption and in many cases indicators will be species-specific.

Existing measures and indicators of welfare tend to fit better with one theory of well-

being or another such as cognitive bias and hedonic theories or choice tests and preferences. Some are closer to their own standalone measure or are measures of health that can then be used as a proxy for the welfare of an animal. Here I examine each indicator and discuss its validity as a measure for an animal's subjective well-being. In many instances, there is insufficient evidence to properly evaluate the indicator on one or more of the criteria that should be used to evaluate a measure. As such I will provide my thoughts given my knowledge of the indicator and criteria.

In all cases, the face validity of the indicator is the perspective of the author of this report even if this would ideally be a composite opinion of multiple experts. An additional concern is that the validation of a measure in one species or genus may not generalise to other taxonomies. A full review of all welfare indicators would validate each measure in every species that it will be used. Unfortunately, we have neither sufficient data nor time to complete this.

SELF REPORTS

Self-reports are any method of measuring welfare that uses the animals' ability to communicate their mental states directly to the researcher. These are mostly only available from humans but a few other primates such as Koko the gorilla (Main, 2018).

Although self-reports in humans are an indicator of the human animal's welfare and should be thought of as a measure of animal welfare like any other, self-reports have been included in this report to provide a comparison between a more well-validated indicator in humans and other indicators.

THE SATISFACTION WITH LIFE SCALE (SWLS)

OVERVIEW OF THE INDICATOR

The SWLS is a 5-item questionnaire to measure subjective well-being developed by Diener et al. (1985) to assess the cognitive judgment component of subjective well-being.

Each item is a statement describing one subjective view of one's life. The overall score is calculated by summing the score for each item. The score for each statement expresses the individual's agreement with the statement. Each item is scored from 1 to 7 on a Likert scale, so the possible range of scores on the questionnaire is from 5 (low satisfaction) to 35 (high satisfaction). The medium score is 20 which expresses the 'neutral' point on the scale.

The full questionnaire is available here.

RELIABILITY	Strength of Indicator: High	
	Strength of Evidence: Strong	

The SWLS scale has repeatedly been shown to be reliable through test-retest reliability.

As shown in the table below, test-retest reliability decreases over time. This is as one would expect where more changes in life circumstances occur over time and change the underlying construct. Test-retest reliability has an average of approximately 0.84 for a period of less than one month which is sufficiently high for its use as a measure.

STUDY	TEST-RETEST CORRELATION	TEMPORAL INTERVAL
Navrátil (2006)	0.90	1 week
Rosengren et al. (2015)	0.78	2 weeks
Pavot et al. (1991)	0.84	2 weeks
Abdallah (1998)	0.83	4 weeks
Pavot et al. (1991)	0.84	1 month
Blais et al. (1989)	0.64	2 months
Diener et al. (1985)	0.82	2 months
Yardley and Rice (1991)	0.50	10 weeks
Magnus et al. (1992)	0.54	4 years

The other measure of reliability that the SWLS is evaluated under is internal consistency. The SWLS has what is usually considered an acceptable to good Cronbach's α value of between 0.74 and 0.85. This is true in samples with different conditions and across cultures. When factor analysis was used, it demonstrated an unidimensional structure, confirming that the SWLS is measuring one underlying construct.

STUDY	CRONBACH'S α	SUGGESTS UNIDIMENSIONAL FACTOR STRUCTURE?
Neto (1993)	0.78	"all items had high factor loadings on a single common factor"
Vassar (2007)	0.78	NA
Gouveia et al. (2008)	0.81	"the results confirmed the single factorial structure"
Beuningen (2008)	0.85	"the five SWLS indicators combine into one factor."

Aishvarya et al. (2014)	0.86	“all items loaded strongly on one-factor solution”
Sagar and Karim (2014)	0.74	“exploratory factor analysis (EFA) (...) identified a single factor structure for SWLS with 5 items.”

FACE VALIDITY

Strength of Indicator: High

Strength of Evidence: Medium

Although social scientists don’t usually use face validity questions when assessing the SWLS, it seems intuitive that the questions asked in the SWLS are attempting to measure the goodness of someone’s life. The only exception to this is the fifth item: “If I could live my life over, I would change almost nothing” which has repeatedly been found to have much lower correlations with the rest of the items and lower factor loadings. This may be due to the more past-facing nature of the question when compared to the other questions.

Face validity seems to hold even when translated into other languages like Bengali. The percentage of respondents who answered “yes” when asked whether the measure is readable, logical, clear, comprehensive, answerable, and had an acceptable style and format was 96.3%, 97.5%, 96.9%, 90%, 81.3%, and 95% respectively (Sagar and Karim, 2014).

CONTENT VALIDITY

Strength of Indicator: Moderate

Strength of Evidence: Medium

The SWLS seeks to capture the cognitive judgment portion of well-being, but not positive and negative affect, or eudaimonic (Niemiec, 2014) assessment of one's life. As a measure of the overall goodness of an individual's life, it only captures some of what we care about. If it could be augmented with either the positive and negative affect scale (PANAS) or the experience sampling method (ESM), it would better capture human well-being.

As it only captures one part of the tripartite model of well-being, its content validity depends on how much weight cognitive judgments have in the assessment of overall well-being. It will also appear more valid to preference utilitarians, as it is an indicator of one's preference for this life compared to others.

CRITERION VALIDITY	Strength of Indicator: High
	Strength of Evidence: Strong
	Cheung and Lucas (2014) showed that the single-item life satisfaction and SWLS were correlated with "theoretically relevant variables, such as demographics, subjective health, domain satisfaction, and affect." This was also found by Jovanović (2016) in adolescents, and by Atroszko et al. (2017) in university students. There are only a few variables correlating in an unexpected direction, which is indicative of high criterion validity.
INTERPERSONAL COMPARISONS OF UTILITY	Strength of Indicator: High
	Strength of Evidence: Weak
	Although harder to assess, I think we can assume with moderate certainty that subjective reports of life satisfaction are valid for ICUs. Plant (2018) examines several possible problems with metrics that depend on subjective reports. This includes utility monsters who are individuals experiencing much greater ranges of utility, or language monsters who experience the same utility range but report different levels of welfare than is accurate. Plant argues that in practice, such monsters are unlikely to occur due to shared language and physiology. Evidence that the same regions light up in the brain across individuals experiencing the same emotions (Davidson and Schuyler, 2015; Kringelbach, 2010) supports Plant's argument. However, these studies only map brain regions to subject reports - they do not provide full knowledge of other minds. To prove that ICUs are not a problem, we would have to directly measure states of consciousness, something that may not even be possible. In the interim, Plant's solution seems adequate for the problems with ICU.

PREFERENCES TESTING

As discussed above, the desire theory of welfare ascribes the satisfaction of an individuals preferences as the focus of their welfare. Therefore to measure this we must determine what an individual's preferences are. As most non-human animals are unable to communicate with us it is more difficult to ascertain what they want. Even in humans where we are able to ask, it's common to find a discrepancy between their stated and revealed preferences. Animal welfare scientists have devised a variety of techniques to try to measure an animal's desire to engage in a certain activity. These can show either positive or negative motivation depending on whether approach or avoided behaviour is exhibited (Kirkden and Pajor, 2006). The two main methods used to assess this fall into two categories, choice tests and operant tests.

Regardless of the method, there are a few overarching practical concerns with preference measurement in non-human animals. Fraser (1997) outlines conditions that can invalidate the findings of such tests. The first is that the experiments accurately reflect the animals' preferences. An animal's preference will vary depending on its age, experience and numerous external factors. So preference experiments need to be comprehensive enough to identify relevant sources of variation. Fraser raises additional concerns that an animal's familiarity with a particular good or environment may affect the results. This can be seen in Dawkins' study of hens preference for housing systems (Dawkins, 1977). Here, he found that the first choice of hens was related to the environment in which they had been living but hens raised in battery cages shifted to the outside run as the trial proceeded.

An additional problem with preference testing comes from variation between individuals. Each individual has their own preferences and although there will be some commonalities between members of the same species their preferences are not guaranteed to be uniform. If we assume that their preferences are uniform we may mistakenly interpret disagreement as a preference for X with some error. This problem is most likely to be encountered on less impactful or positive goods such as forms of environmental enrichment. For example, in a hypothetical experiment 75% of pigs may show a preference for blue balls over red balls for environmental enrichment. Despite this, we should be hesitant to conclude that all pigs prefer blue balls since it's possible that the other 25% do prefer red balls rather than some pigs accidentally picking the less preferred ball.

The main restriction to measuring animal preferences is that we are only able to test preferences that an animal can reasonably understand. Which makes it difficult to measure a choice that falls outside of the animals sensory, cognitive and affective capacities, or if animals are required to choose between short- and long-term benefits (Fraser, 1997).

CHOICE TESTS

OVERVIEW OF THE INDICATOR

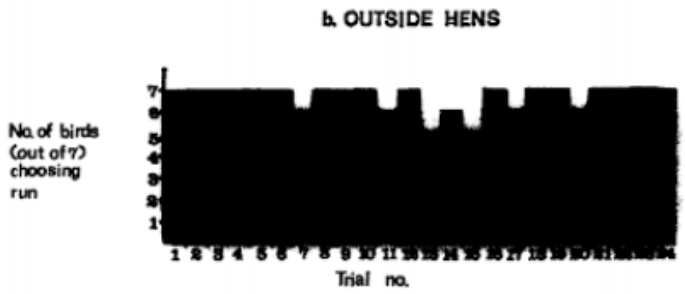
Choice tests offer animals the options to choose between two or more different environments or resources (Kirkden and Pajor, 2006). This can give a clearer sense of which of two goods an animal preferences, based on the probability of them choosing each good.

RELIABILITY

Strength of Indicator: High

Strength of Evidence: Weak

As shown by Dawkins the reliability of such a test relies on the length of the trial (Dawkins, 1977). If the trial is of insufficient length, the animal will not have enough time to learn what their preferences are. This will vary depending on the cognitive sophistication of the animal and on the strength of the cues for the choice. Rainbow trout for example were able to learn to avoid a frightening stimulus by swimming through a doorway to another chamber after a few exposures but took many days to learn to associate this with a conditional stimulus (the illumination of a light) (Yue, Moccia and Duncan, 2004). When the animal has learned their preference and the conditional stimulus is strong, then the reliability of this preference over repeated trials seems high. However, no formal evaluations of animal preference through test-retest reliability were found.



(Dawkins, 1977)

FACE VALIDITY

Strength of Indicator: High

Strength of Evidence: Conjecture

The face validity of choice tests as a measure of the animal's preference is high as we are evidently testing the animal's revealed preferences.

CONTENT VALIDITY	Strength of Indicator: Low Strength of Evidence: Conjecture

The results of a choice test can only reveal the ordinal preference of an animal for the limited number of options presented. This means that on their own, choice tests can't help us decide whether a preference is significant to an animal or not. Even with multiple tests, it is not clear that an interval scale of the degree of preference can be found (Rushen, 1986). An individual's choice between tea or coffee or between social isolation or companionship would appear the same even if the effect on an individual's welfare may be much greater with the latter.

CRITERION VALIDITY	Strength of Indicator: High Strength of Evidence: Medium

Preference is seen as a valid indicator of welfare in humans and aligns with numerous positive outcomes. Beyond the goods that are necessary for survival, most people want to earn more money and have good social relationships - both of which affect life satisfaction (McGuire, 2020; Barger, Donoho and Wayment, 2009). There are numerous other examples of this that we are aware of from our own experience but preferences and well-being can sometimes diverge. Many individuals will continue to strive for a higher income even beyond the point of diminishing returns for well-being.

In other animals, the kinds of preferences we are measuring are much simpler. Their inability to think about their long term preferences means we can only assess immediate choice. Nicol et al. validated hens' choice of three environmental conditions by examining their physiological response in three different environments given their observed preferences for each (Nicol et al., 2009). They found that preferred environments were associated with lower body temperature, blood glucose, heterophil:lymphocyte ratio and response to novelty, as well as greater feed digestibility and self-grooming. However, other indicators that are believed to be valid indicators of welfare were not associated with their choice.

INTERPERSONAL COMPARISONS OF UTILITY

Strength of Indicator: High

Strength of Evidence: Conjecture

Choice tests allow us to clearly compare the preferences of two individuals or groups.

OPERANT TESTS

OVERVIEW OF THE INDICATOR

Operant Tests are used to measure an animal's 'willingness to pay', a common currency. The currency is usually learned behaviours such as pushing on a heavy door, pulling a loop or pecking a button (Amdam, 2011). This allows researchers to measure an animal's relative preference for multiple goods such as social interaction or food.

RELIABILITY	Strength of Indicator: Moderate	
	Strength of Evidence: Conjecture	
	It is theoretically possible to assess the reliability of an individual's willingness to pay for a good through such trials but no existing studies were found. Although it is worth noting that one would expect an individual's willingness to pay to vary dramatically over short periods of time due to differences in mood and level of satiety for that good. This doesn't indicate that operant tests are unreliable, only that the underlying preference is in flux so we should expect lower reliability when tested. A priori, I would expect an individual's revealed preferences to map reliably onto their true preferences at least for short-term rewards.
FACE VALIDITY	Strength of Indicator: High	
	Strength of Evidence: Conjecture	
	As with choice tests, operant tests have high face validity as a measure of an individual's preference.
CONTENT VALIDITY	Strength of Indicator: High	
	Strength of Evidence: Conjecture	
	Operant tests can, in principle, provide information about both the direction of the animal's preference and the strength of these preferences. There are only practical or ethical limitations to their use. Individuals may not understand the test, may be unable to understand that work can pay off in the future and ethically complicated to test their willingness to pay to escape pain.

CRITERION VALIDITY	Strength of Indicator: High Strength of Evidence: Weak Many animals have demonstrated a willingness to pay for food that increases as the time since the last meal increases. Comparing an animal's willingness to pay for food after different lengths of time to other goods, such as social contact, has become a standard method for assessing the strength of their preference (Akre, Bakken and Hovland, 2009). There are a few instances in both human and non-human animals where an animal's preference for a good diverges from its welfare, namely addiction. These are instances where an animal's preference diverges from what we as observers think is best for their welfare. Another potential problem is contrafreeloading, which is when an individual will choose to work for food rather than take food that is offered for free. This suggests that some animals are motivated to 'work' even if the payoff is the same which may artificially inflate their willingness to pay.
INTERPERSONAL COMPARISONS OF UTILITY	Strength of Indicator: Moderate Strength of Evidence: Conjecture Comparing the preference strength between individuals is more difficult as an individual's ability to pay a common currency will vary depending on their characteristics, such as increased physical strength. If all else is held constant, operant tests are a good method to compare the strength of individuals' preferences.

COGNITIVE BIAS TESTING

OVERVIEW OF THE INDICATOR

Research into human psychology has shown that someone's affective state influences their cognitive function. Those suffering from or who have had depression can suffer from negatively biased information processing (Gotlib and Krasnoperova, 1998). The recognition of this phenomenon in humans led to the testing of cognitive biases in other animals. This is tested through the 'judgement bias' paradigm in which animals are trained so that one cue predicts a positive event and another cue predicts a less positive/negative event, and are then presented with ambiguous cues (Mendl, Burman, Parker, Paul, 2009). If the findings generalise from humans then we expect that animals in a negative affective state will be more likely to judge these ambiguous cues as predictive of a negative event than animals in a more positive state. This means animals in a more negative affective state are more likely to avoid ambiguous cues than animals in a positive state.

RELIABILITY	Strength of Indicator: Moderate
	Strength of Evidence: Conjecture
	No evidence of the reliability of cognitive bias tests was found during this review. As cognitive biases are built up over a longer period of time, one would expect animals to display these biases during repeated testing. However, over time the animal may learn that the ambiguous cue is not associated with a punishment, though this is unlikely to occur over the short-term. So a priori I would expect cognitive bias testing to be reliable. If it is not, this would indicate some noise during its measurement.
FACE VALIDITY	Strength of Indicator: Moderate
	Strength of Evidence: Conjecture
	These biases develop as a protective mechanism so are face valid as indicators of which events and goods an individual prefers. However, measuring the cognitive bias of individuals does not seem to intuitively track onto affect. Given that, a priori one might not expect those with a stronger negative bias to have a lower affect after the period when the association is being built. If these biases are an indication of negative thought patterns, then these will negatively affect one's perception of the world and the affect.

CONTENT VALIDITY	Strength of Indicator: High	
	Strength of Evidence: Conjecture	
	Both optimistic and pessimistic biases have been demonstrated depending on the affective state of the individual. Therefore, cognitive bias seems to capture the balance between both positive and negative affect. Theoretically, the strength of bias is an indicator of the degree of imbalance between these, even if in practice this may be harder to evaluate.	
CRITERION VALIDITY	Strength of Indicator: High	
	Strength of Evidence: Strong	
	Pessimistic cognitive bias has been repeatedly correlated with negative emotional states in humans (Mendl , Burman, Parker, Paul, 2009; Eysenck, Mogg, May, Richards, & Mathews, 1991; Beard, 2009; Krantz & Hammen, 1979). This same phenomenon has been documented in other animals with pessimistic and optimistic biases, observed in the expected direction. These include mammals such as rats (Brydges, 2011; Enkel et al., 2009) and sheep (Doyle, 2010), as well as animals less closely related to us such as zebrafish (Wojand, 2015) and honeybees (Bateson *et al*., 2011).	
INTERPERSONAL COMPARISONS OF UTILITY	Strength of Indicator: Moderate	
	Strength of Evidence: Weak	
	The strength and direction of an individual's cognitive bias will vary depending on their personality or culture (Chang, Asakawa, & Sanna, 2001), both in humans (Marshall et al. 1992) and non-human animals such as pigs. Asher et al. (2016) found that pigs with a more proactive personality were more likely to respond optimistically to unrewarded ambiguous stimuli than reactive pigs. This makes it harder to compare two individuals' levels of bias to determine their effect due to variations in their personalities. This makes it necessary to use a larger sample to ensure that the cognitive biases of the animals are due to circumstance rather than personality.	

PRIMARY BIOLOGICAL INDICATORS

OVERVIEW OF THE INDICATOR

Primary biological indicators include any primary psychological responses within an animal. These occur when neuroendocrine cells receive neuronal input and then release hormones into the blood. This can occur because of stressful stimuli or because of positive experiences. Stress indicators include adrenal activity, cortisol, and norepinephrine. Hormones associated with positive experiences include dopamine, serotonin, oxytocin, and endorphins.

The concentration of these hormones can be measured using invasive sampling methods such as blood tests, or other non-invasive methods such as saliva, excreta, milk, hair/feathers, and eggs (Palme, 2012). What each indicator tells us about an individual's subjective experience depends on the item of interest and the methods of collection. For example, catecholamines and glucocorticoids are released within seconds to minutes after a stressor and then quickly metabolised and excreted via urine and faeces (Palme, 2012). This makes blood concentration of these hormones a poor long-term indicator of stressors and excreta a more stable one.

In an extensive review, each of these indicators would be examined in their own right but we will not go into such depth.

RELIABILITY	Strength of Indicator: Moderate
	Strength of Evidence: Weak
	The sampling method for each hormone is reliable in that each measurement is a reliable measurement of the hormone present in that sample. The reliability of such measures for an overall assessment of the concentration of a hormone in an animal varies depending on where the sample is taken from and the hormone. For several methods of sampling saliva, there are additional accuracy concerns as hormone levels vary over short periods of time due to the animal's circadian rhythm (Lane, 2006). One exception to this is faecal measurements of glucocorticoids metabolites as they are not affected by the time of day of faecal deposit (Lane, 2006). This means that accurate comparisons of hormone levels requires an understanding of the base-line variation in the animal (Cavigelli et al., 2005).

	The most widely used primary indicator of welfare is cortisol. Short et al. compared the reliability of cortisol measures in hair, saliva, and urine in humans (Short et al., 2016). They found a strong association between hair samples and the previous 30-day average salivary cortisol but did not find a significant relationship between hair and integrated urine samples. They also found variation in the test-retest reliability of sampling methods. Hair had the highest reliability (month-to-month: $r = 0.84$, $p < 0.001$), then urine C (week-to-week: r's between 0.59 and 0.68, ps < 0.05), whilst saliva was the least reliable (week-to-week: r's between 0.38 and 0.61, p's between 0.13 and 0.01) (Short et al., 2016). We suspect that the observed variation in test-retest reliability between sampling methods is more indicative of a greater sensitivity to the variation in the underlying cortisol levels over time, rather than lower reliability. This can be seen in a sample of cows studied by Verkerk et al. (1998) They injected an adrenocorticotropic hormone (ACTH) into three groups of lactating cows at different time intervals before milking. They found that mean cortisol concentrations in their milk were similar for the group injected four hours before and control hour groups (1.2 and 0.5 ng ml−1, respectively), but higher in the two hour (2.4ng ml−1), and one hour groups (11.7 ng ml−1; $P < 0.001$). This highlights a concern that some sampling methods are only reliable indicators of short-term spikes in hormones and can therefore only be used to measure short-term acute stressors such as handling.
FACE VALIDITY	Strength of Indicator: Moderate Strength of Evidence: Conjecture If we assume that physical states are directly related to subjective states, then a collection of physiological indicators could be a valid way to assess welfare. It is not evident a priori which hormones are associated with well-being and which subjective states each hormone maps onto.
CONTENT VALIDITY	Strength of Indicator: Moderate Strength of Evidence: Weak

	Each hormone is an indicator of one aspect of well-being so as an isolated indicator has low content validity. In humans, some hormones such as cortisol track negative affect (Buchanan, al'Absi and Lovallo, 1999) whilst others such as serotonin are only associated with positive affect (Williams et al., 2006). A combination of these indicators could be used to gather information about many possible subjective states, but even with this information quantitatively measuring the affect balance would not be possible.
CRITERION VALIDITY	Strength of Indicator: Moderate
	Strength of Evidence: Medium
	As mentioned above, in humans cortisol increases with negative affect (Buchanan, al'Absi and Lovallo, 1999) and serotonin is correlated with positive and not negative affect (Williams et al., 2006). Saliva cortisol samples have also been mildly negatively associated with global life satisfaction measures (Smyth et al., 2017). Cortisol is also associated with the expected variable in other animals. In pigs, hair cortisol increases with tail lesions or lameness (Carroll et al., 2018) and plasma-free cortisol is higher in farrowing crates compared to pen systems (Cronin et al., 1991). Chickens and humans show elevated cortisol levels when exposed to loud noises (Lee, Kim and Lee, 2003; (Jafari et al., 2019). Other acute stressors such as net handling increase whole-body cortisol levels in zebrafish up to one hour post-stressor (Ramsay et al., 2009). Even though these hormones tend to track onto the expected experiences and other validated constructs, such as positive and negative affect schedule (PANAS), in some circumstances they can diverge.

West et al. (2004) compared the effects of three 90-minute classes: dance, yoga, or a lecture on PANAS and saliva cortisol levels. Dance and yoga reduced negative affect. However, cortisol increased in African dance and decreased in Hatha yoga. West notes that this shows that ‘even when these interventions produce similar positive psychological effects, the effects may be very different on physiological stress processes’.

This is also seen in other animals. Colborn et al. (1991) found that stallions showed similar cortisol responses whether they were permitted to mate with a mare, had acute physical exercise, or were restrained. These experiences will have different effects on the animals’ welfare but by using cortisol alone we cannot observe any difference. Thus, it seems that individual physiological responses cannot reliably be taken in isolation as they can track both arousal and affect.

INTERPERSONAL COMPARISONS OF UTILITY

Strength of Indicator: High

Strength of Evidence: Weak

The hormone concentration of a sample can be easily compared between individuals but what this implies for their subjective experience is less clear. Individuals’ baseline concentration may vary so that given the same circumstances they have different cortisol levels. This could be an indication that one individual has a naturally higher stress response, but it could be that these different levels are experienced the same subjectively. Within a species, we can assume that our shared physiology will result in a similar subjective experience (with the possible complication of minor sex differences in sexually dimorphic species).

When comparing between individuals with more genetic variation, this problem is much greater. Lane argues that as basal levels of GCs are seen to vary hugely between species, it precludes any direct point comparisons between species (Lane, 2006). Instead, we need an understanding of the basal levels and the relative change during an individual's response.

SECONDARY BIOLOGICAL INDICATORS

OVERVIEW OF THE INDICATOR

Secondary biological indicators occur as a consequence of the release of primary biological indicators. Although they are still indicators of the subjective experience of the animal, they are further down the causal chain from their welfare. These indicators include blood metabolite concentrations such as glucose and lactate, gastrointestinal activity (Moberg and Mench, 2000), metabolism, heart rate and heart rate variability (von Borell et al., 2007), and respiratory functions (Barton, 2002). These are all measured through either physical samples or various portable equipment available for non-invasive methods, such as electrocardiography.

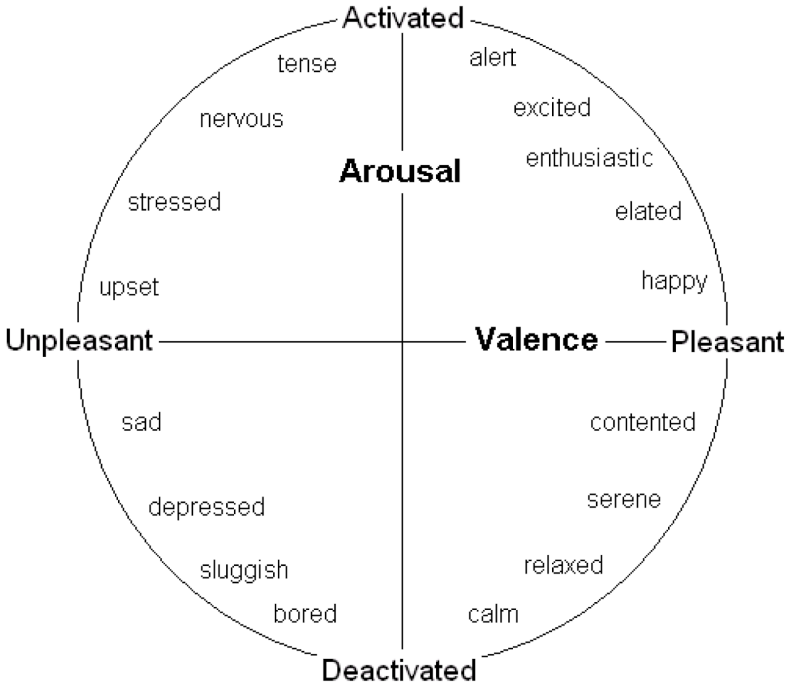
RELIABILITY	Strength of Indicator: Moderate	
	Strength of Evidence: Medium	
	As with primary biological indicators, invasive methods produce reliable measurements of the presence of an indicator in a sample. External measures such as heart rate, heart rate variability, and respiration rate are also highly reliable measures (Guijt, Sluiter and Frings-Dresen, 2007; Drost, 2011), dependent on the device used to measure them. Unlike primary indicators reliability does not depend on the sample used as each indicator is measured in the sample of interest. Take heart rate where we can take measures reliably from pulse rather than approximating heart rate through pulse as cortisol in an individual's hair relates to the cortisol in their blood.	
FACE VALIDITY	Strength of Indicator: Low	
	Strength of Evidence: Conjecture	
	Most indicators that fall into this category don't obviously track the valence of the subjective experience of an individual. Many will be weak indicators of intensity if changes are measured over time. Generally, these indicators seem weak because they are too far down the causal chain for us to reliably make inferences back to an individual's welfare.	

CONTENT VALIDITY

Strength of Indicator: Low

Strength of Evidence: Conjecture

As many of these indicators have low face validity, it is not obvious which aspects of well-being they track onto. Many respond to both positive and negative experiences (see below) in the same way and thus don't properly track any single aspect of well-being. Instead, they seem to indicate the degree of arousal associated with an experience rather than its valence.



(Trimmer et al., 2013)

CRITERION VALIDITY

Strength of Indicator: Low

Strength of Evidence: Medium

Several secondary biological indicators are associated with both positive and negative valence experiences. For example, in humans increased heart rate has been associated with both excitement (Wulfert et al., 2005) and fear (Sartory, Rachman and Grey, 1977).

In other animals, secondary biological indicators have been associated with other variables suspected to create high arousal states. Captive European starlings' heart rate increases when exposed to various stressors (Nephew, Kahn and Romero, 2003). Stressed fish show the expected increase in glucose and lactate levels (Carragher and Rees, 1994). In Jersey calves, non-optimal temperatures result in raised respiration and heart rate (Kristensen T N, 2006). Lambs transported in the winter have higher glucose and cortisol levels (Miranda-de la Lama et al., 2010).

Many of these changes can be thought of as the body of the animal's attempt to adapt or prepare them to deal with a stressor. This raises issues with the validity of many of these measures for measuring welfare as extreme stressors can surpass this limit. An example of this can be seen in fish where low oxygen levels can result in a slowed heart rate (Brijs et al., 2018) even if this places stress on the fish.

Many of these indicators naturally increase due to other stimuli or activities. Heart rate, lactate, and respiration increase during exercise and glucose rises sharply after consuming carbohydrate-rich meals (Steffens, 1969).

INTERPERSONAL COMPARISON OF UTILITY

Strength of Indicator: High

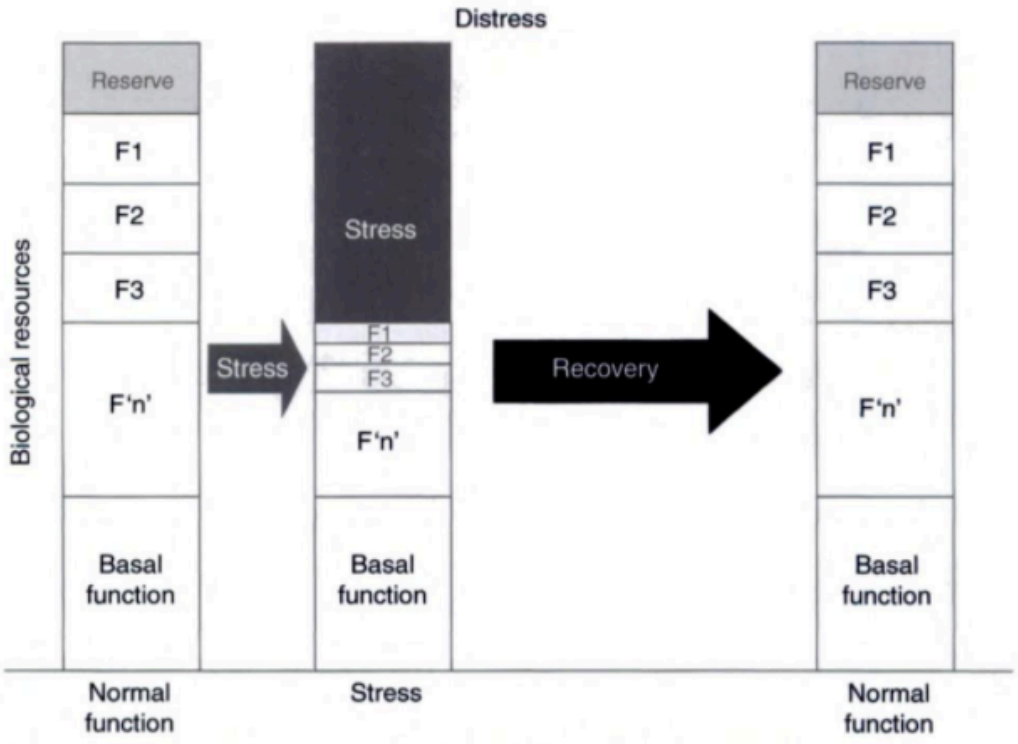
Strength of Evidence: Conjecture

As measures of welfare, it’s not clear that individuals will respond similarly given the same stressor or based on their welfare. Different individuals will have different capabilities when responding to stressors so the same increase in an indicator may not imply the same change in welfare. This is especially true when comparing across species as there is a large variation in heart rate (Comparing the heart rates of animals and human beings, no date) and this is also likely for other indicators.

TERTIARY BIOLOGICAL INDICATORS

OVERVIEW OF THE INDICATOR

Tertiary biological indicators occur when a stressor is continuous enough that it exceeds the body's attempts to cope. These refer to aspects of the whole performance of an animal. This is explained by Moberg's hypothetical model of the expenditure of an individual's biological resources as seen below (Moberg and Mench, 2000). When the stressor exceeds the body's reserve capacity, it becomes 'distressed' and shifts the metabolism away from other functions. We evolved this capacity to cope with extreme short-term stressors such as predation where diverting resources away from other functions could prevent our death.



(Moberg and Mench, 2000)

Some examples of this are reductions in feed intake, ovulation, growth rates, impaired reproduction, milk production, wool and egg production, and immunosuppression. Many of these indicators are also measures of productivity in farmed animals.

Most of these indicators can be measured by weight or volume. The more complicated indicators such as immunosuppression are measured by alterations of lymphocyte or macrophage function. These are tested through the presence of various white blood cells in physical samples.

RELIABILITY	Strength of Indicator: High	
	Strength of Evidence: Medium	
	Most tertiary biological indicators can be measured non-invasively but have a great deal of natural variation. The measurements themselves are reliable as feed intake, growth rates, milk, wool and egg production can all be measured by volume or weight. These will be subject to some measurement error but this is unlikely to be significant. Some indicators, such as growth rates, can increase their reliability through multiwave testing (Willett, 1989), whilst immunosuppression measures are mostly reliable. In one study, errors due to operator and random error were 3.12% and 7.8% respectively (Kristensen T N, 2006).	
FACE VALIDITY	Strength of Indicator: Low	
	Strength of Evidence: Conjecture	
	Many of these measures do not have a strong a priori relationship with an animal's emotional state. Production measures in particular could be completely disconnected from welfare. Others such as feed intake could go either way as both low and high feed intake could be an indicator of emotional distress. Immunosuppression seems the most face valid but it is still only a proxy for welfare rather than a direct measurement. This is also complicated by the fact that immune function is both an indicator and a condition for future welfare.	
CONTENT VALIDITY	Strength of Indicator: Low	
	Strength of Evidence: Weak	
	Here it is evident that each measure does not capture the full underlying construct of the animal's well-being. All tertiary biological indicators only provide information about negative valenced experiences, particularly stress. In addition, as Moberg's model shows, many of these indicators are only evidence of major stressors (Moberg and Mench, 2000) and are therefore unable to capture low arousal experiences.	

CRITERION VALIDITY

Strength of Indicator: Low

Strength of Evidence: Medium

Tertiary biological indicators are associated with numerous factors, including factors we expect are associated with an individual's welfare.

In pigs, unpleasant handling has been associated with a decreased growth rate and higher cortisol (Hemsworth, Barnett and Hansen, 1981). In ruminants, heat stress has been shown to lower milk production and decrease growth rate for cattle and lambs, but has little effect on wool production (Morrison, 1983). McKenzie et al. (2012) examined rainbow trout and found that higher stocking densities (~ 75 and ~ 100 kg m− 3) were associated with lower growth rates. However, there was no evidence of a neuroendocrine stress response when compared to a low-density control.

There are numerous immunosuppressive agents including pathogens, drugs, malnutrition, weaning, and stress (Muneer et al., 1988), though, the relationship between stress and immune function is weak. Segerstrom and Miller (2004) found that self-reported stress was poorly associated with immune function, although there was little research into this association. More recent research from Hameed (2018) found a negative association between Immunoglobulin A and perceived stress but a positive association between Secretory Immunoglobulin and depression and loneliness.

Many things that affect tertiary indicators also affect welfare but many of these indicators rely more on this correlational evidence. Immune response can be influenced by stress or numerous other factors, as can growth rate and production indicators.

INTERPERSONAL COMPARISONS OF UTILITY

Strength of Indicator: Moderate

Strength of Evidence: Conjecture

It is possible to compare growth rates and immune function across individuals or species but what this indicates in regards to welfare is unclear. There are large variations on many of these indicators between individuals of the same species. In one experiment, egg production can vary between 8 and 25 eggs in a 28 day period (Carlander, Wilhelmson and Larsson, 2001). Variation in growth rates between individuals are evident in many species, including humans (Eichorn, 1968), and can be particularly pronounced between sexes.

Again these issues are even more pronounced between species or different breeds. Comparing the growth rates of a chicken to a trout will tell us nothing about their comparative welfare. Using these indicators in relation to their maximum under ideal conditions somewhat solves this problem but this will still be a poor indicator of their relative welfare.

BIOLOGICAL MARKERS OF AGEING

OVERVIEW OF THE INDICATOR

Biological markers of ageing are a kind of tertiary biological indicator but they are worth examining separately as their measurement may allow us to measure the cumulative welfare of an individual. Markers of ageing are measures of an individual's biological age - the degree of age-related change/deterioration in appearance, health, or functionality - when compared to their chronological age. A wide variety of biological measures can be used such as ‘telomere length and attrition rate, DNA methylation patterns, gene expression profiles, changes in neuroanatomy, proteomic and metabolomic changes, and various composites of clinically relevant symptoms’(Bradshaw, 2019). If an individual's biological age is high relative to their chronological age then this is an indicator that their welfare is poor (Bradshaw, 2019). The cheapest and therefore most common method for assessing biological age is telomere length so this will be the main focus here.

RELIABILITY	Strength of Indicator: High	
	Strength of Evidence: Medium	
	Telomere length has been shown to be reliable through repeated testing in a human sample. Kim et al. (2011) found that telomere length from 7 blood samples taken over a 7 month period from 27 non-pregnant adult women (aged 35 to 74 years) did not differ significantly from the mean of the samples. The method of testing and storage can affect the reliability of tests. Eastwood et al. (2018) found that inter−extraction repeatability was 50% lower for samples stored in a buffer solution. Other biological markers of ageing, such as DNA methylation where DNA sourced from lymphoblastoid cell lines, showed distinct patterns when compared to both blood and saliva (Thompson et al., 2012).	
FACE VALIDITY	Strength of Indicator: Moderate	
	Strength of Evidence: Weak	
	Bradshaw (2019) outlines the theoretical reasons to expect the rate of biological ageing to correlate with the cumulative affective experience of an individual (Thompson et al., 2012).	

	Through an evolutionary lens, experiences are used to motivate evolutionarily advantageous actions so we should therefore expect the valence of an experience to correlate with its effect on an animal's fitness. One way this can manifest is through damaging or repairing the body. Therefore, we should expect experiences that positively or negatively physiological affect an individual to track onto their valence. Since ageing is a progressive accumulation of physiological damage, we should expect the valence of an experience to also generally correlate with its effect on biological age.
CONTENT VALIDITY	Strength of Indicator: Moderate
	Strength of Evidence: Conjecture
	Theoretically, biological ageing captures the cumulative affect balance of an individual's experiences. However, there are some aspects that it is unable to capture, such as end of life experiences (Bradshaw, 2019). It is also unclear in practice whether any experience regardless of intensity will affect biological ageing.
CRITERION VALIDITY	Strength of Indicator: High
	Strength of Evidence: Medium
	The relationship between biological markers of ageing and the welfare of an individual is well validated in humans, though the strength of this relationship is often small. A meta-analysis from Pepper et al. (2018) found that there was a weak association between telomere variables and exposure to a variety of stressful events, including physical diseases, environmental hazards, and psychiatric illness ($r = –0.09$, 95% CI –0.13 to –0.05). Schutte et al. (2015) had similar findings through a meta-analysis of the association between depression and telomere length ($r = –.12$, $P < .001$). However, there is some evidence that a healthy lifestyle can mitigate the effect of major life stressors on telomere length (Puterman et al., 2015). This is evidence of the protective effect of a good lifestyle physiologically, though it is unclear whether the same effect occurs psychologically.

These indicators have been less well-validated in other species but there is some evidence from various taxa. These include a decrease in telomere length from reproductive stress in mice (Kotrschal, Ilmonen and Penn, 2007), high stocking density in broiler chickens (Beloor et al., 2010) and salmon with an artificially enhanced growth rate (Pauliny et al., 2015). For most species further validation is needed against other welfare indicators.

INTERPERSONAL COMPARISONS OF UTILITY

Strength of Indicator: Moderate

Strength of Evidence: Weak

Due to natural genetic variation, individuals naturally age at different rates. One study in humans found that 57% of the variation in biological age could be explained by genetic factors (Karasik et al., 2004). This makes it difficult to use the rate of biological ageing of two individuals as a direct comparison of their welfare. However, in larger samples where this natural variation can be accounted for, intra-species group comparisons of welfare are possible. Nevertheless, even in such a case, it is still difficult to compare individuals at different life stages (Bradshaw, 2019).

When looking to compare across species this problem is even greater. First, there is natural variation in rates of ageing, time spent as an adolescent and overall lifespan. It's also not clear that stressors of equal significance to two species will have the same effect on the relative speed of biological ageing compared to controls. This is in part determined by the marker that is used, with telomere attrition commonly viewed as the most cross applicable (Bateson, 2016).

In species particularly distinct from humans, the validity of the rate of biological ageing as a measure of welfare is questionable (Bradshaw, 2019). For example, in cases where their pattern of ageing is different from humans, such as the desert tortoise (Gopherus agassizii) whose mortality decreases as it gets older (Gewin, 2013).

PHYSICAL HEALTH

OVERVIEW OF THE INDICATOR

The health and biological function of an animal is seen as a prerequisite to good welfare. As discussed in the first section of this report, biological function is also commonly viewed as one part of the definition of animal welfare. Health can be measured in a variety of ways and is typically the most commonly used indicator for farmed animals. Common measures of health include mortality, disease rates, and injury rates. For example, mortality and rates of disease are widely monitored on farms in large part due to their connection to productivity.

Health indicators can also include tertiary biological indicators such as immune function, growth, and reproduction. However, I will only be examining broader physical rather than physiological indicators in this section.

RELIABILITY	Strength of Indicator: High
	Strength of Evidence: Weak

Mortality can be measured easily and accurately through on-farm monitoring of losses. This level of monitoring can vary from a basic recording of losses to detailed records of the date, age, weight and cause of death for each individual. Although we found no formal assessments of the reliability of mortality, we expect it to be a reliable measure because of the mostly clear distinction between the alive and dead.

Disease rates are commonly assessed through the causes of mortality determined from a necropsy. This is assessed with some animals and farms and not others. As above, no formal assessments of inter-rater reliability of cause of death have been done. A priori, we expect this to be fairly reliable but that there will be some level of human error. Part et al. (2016) are also fairly confident that disease prevalence rates will be fairly reliable as they are performed by official veterinarians and there is a strong food safety incentive for accurate diagnostics.

	Injury and animal condition grading are assessed by external observers. The reliability of these measures has been well studied for a variety of animals and injuries. In chickens, these include keel bone fractures which have moderate inter-observer reliability (r=0.44) and high accuracy (Petrik, Guerin and Widowski, 2013), a feather scoring system with very high inter-observer reliability (r=0.88)(Decina et al., 2019), and foot pad dermatitis scoring which had an average PABAK value of 0.88 (Piller et al., 2020). Similar findings in cows for lameness (Willen, 2010) and body condition score (Morin et al., 2020), and in pigs for tail lesions, body lesions, and lameness (Mullan et al. 2011) show these measures are reliable. Phythian et al. (2012) examined the reliability of multiple indicators of sheep welfare and found that inter-observer reliability was high for lameness, wool loss, cleanliness of the breech area, and ventral abdomen. The other behavioural indicators examined occurred too infrequently in the sample to be reliable.
FACE VALIDITY	Strength of Indicator: Low
	Strength of Evidence: Conjecture
	Good health may correlate with many factors included in our definition of welfare. Poor health will often lead to poor welfare but health in and of itself is not a measure of welfare.
CONTENT VALIDITY	Strength of Indicator: Low
	Strength of Evidence: Conjecture
	Good physical health is an important contributor to the welfare of an animal but does not capture the full construct. Segner et al. (2012) writes ‘the fact that an animal is healthy does not necessarily mean that it has a good welfare status. Thus, welfare is the broader, more overarching concept than the health concept.’ In a practical setting, particularly with indicators such as mortality, measures of health only serve as evidence of very poor welfare.

CRITERION VALIDITY	Strength of Indicator: Moderate
	Strength of Evidence: Medium

A variety of health conditions and reduction in functioning have been shown to have an effect on human well-being. These include pain, conditions that reduce our ability to function, or severe restrictions on our mobility (Dolan and Metcalfe, 2012). There is some evidence of a relationship between mortality and subjective well-being. A 2017 meta-analysis showed that subjective well-being was a protective factor for mortality (pooled hazard ratio = 0.920) (Martín-María et al., 2017). Although the author notes that additional research would be required to 'establish a cause-effect relationship.'

Although physical health has been shown to track onto well-being the majority of the time, in some cases there can be a large degree of divergence. Dolan and Metcalfe (2012) examine subjective well-being as the dependent variable against dummies for all 5 factors in the EQ5D (a measure of health and functioning). They found no statistically significant effect on either life satisfaction or day effects of those with the condition "some problems walking", a condition that participants were willing to trade off 14.6% of their remaining life to avoid.

In non-human animals, the relationship between health, mortality, well-being and stress has also been documented. For salmon, stressful events compound to increase the risk of mortality when exposed to additional stressors (Järvi, 1989). Housing systems can have a large effect on the distribution of mortality amongst laying flocks (Weeks, Lambton and Williams, 2016). Some measures of health have also been validated against biological indicators. Lame pigs and prolapsed pigs with pain showed higher salivary levels of cortisol compared with healthy pigs (Contreras-Aguilar et al., 2019).

INTERPERSONAL COMPARISONS OF UTILITY

Strength of Indicator: Low

Strength of Evidence: Conjecture

It's possible to compare health states and functioning between individuals but this does not necessarily give us a fair comparison of an individual's welfare. Take the example above of “some problems walking”. Some individuals will value fully functioning walking a lot while others will not. Somebody who enjoys hiking regularly will be affected much more than someone whose main hobby is watching movies. Even with particular diseases or painful injuries, there is variation in each individual's ability to cope.

ABNORMAL BEHAVIOURS

OVERVIEW OF THE INDICATOR

Abnormal behaviours tend to develop in adverse and stressful environments. These behaviours are an attempt for an animal to mentally cope with the stressors. Abnormal behaviours are identified when they meet some or all of the following criteria (Garner, 2005):

- The behaviour is only seen in captivity
- If seen in the wild, the behaviour is performed in inappropriate circumstances or performed excessively
- The behaviour involves self-injury, affects social interactions, or has deleterious consequences on growth or reproduction
- The behaviour is peculiar to a subset of individuals
- The behaviour induces signs of distress in the animal or its companions

Examples of such behaviours include stereotypies (repetitive movements), self-mutilation, tail-biting, feather-pecking, aggressive behaviour, and vacuum activities. The presence and intensity of such behaviours are measured by observers who document the frequency and time spent performing these behaviours. Some behaviours can be documented through the consequences of the behaviours, such as with tail-biting and feather-pecking.

RELIABILITY	Strength of Indicator: High
	Strength of Evidence: Weak
	The reliability of abnormal behaviours as an indicator of welfare falls into two categories: that the same conditions result in the same response and that methods for assessing the severity and prevalence of these behaviours are reliable. The first aspect of this, that the same environmental conditions will result in the appearance of the same abnormal behaviours, is doubtful. This is driven by the fact that abnormal behaviours are a result of an individual's response to psychological damage.

	Garner (2005) notes that some individuals will perform abnormal behaviours while others will not, even when they are of the same strain, sex, and age, experiencing the same housing, husbandry, and handling, and housed in the same cage. Even amongst those animals that perform abnormal behaviour, individuals differ significantly in the severity of the behaviour. Although the performance of abnormal behaviours by an individual is unreliable, across a large enough population we should expect the incidents to be a better indicator. Given abnormal behaviours are present, we then need reliable ways of assessing their severity and frequency. This varies from behaviour to behaviour but the most well studied are behaviours that result in physical damage, such as feather pecking. These can be assessed through body condition scores as with other health indicators. Feather scoring systems have high inter-observer reliability (r=0.88) (Decina et al., 2019) (r=0.89) (Van Zeeland et al., 2013) and high intra-observer reliability (0.9-0.95). Behaviour scoring systems have been used in some animals, including elephants where they had a moderate level of inter-rater reliability of r=0.56 (Yon et al., 2019). However, the reliability of the vast majority of measures of behaviour has not been measured in each species.
FACE VALIDITY	Strength of Indicator: Moderate Strength of Evidence: Conjecture The performance of abnormal behaviour, as defined by Garners (2005) criteria, is somewhat likely to be an indicator of stress or negative welfare. However, in some instances it's possible that the same behaviours could be displayed when the animal is experiencing positive, neutral or negative welfare. Even still, abnormal behaviour appears to be a weak indicator of the current welfare of the animal and is more often considered an indicator of past welfare issues.
CONTENT VALIDITY	Strength of Indicator: Low Strength of Evidence: Conjecture

Most behaviours that fall into this category are exhibited when an individual experiences sufficient stress (either chronic or acute) to cause it to use behaviour as a coping mechanism. There are some instances where behaviours such as stereotypies are associated with positive or neutral welfare, such as thumb sucking in human infants. However, most instances of abnormal behaviour are only associated with negative welfare. The performance of these stereotypies are a coping mechanism for stress, so may be beneficial to an animal's welfare but ultimately indicate the presence of a stressor.

CRITERION VALIDITY

Strength of Indicator: High

Strength of Evidence: Weak

Some abnormal behaviours have been validated in humans against measures of subjective well being. For example, self-harm is associated with higher negative affect (Croyle, 2000; Ana Xavier, 2014). Conditions that result in similar behaviours, such as obsessive compulsive disorder, have also been associated with lower levels of life satisfaction (Norberg et al., 2008). However, some behaviours seem to emerge at certain life stages regardless of welfare, such as rhythmic stereotypies in infants (Thelen, 1979). These can be elicited by a wide variety of contexts including non-alert states, interactions with the caregiver and other persons, feeding situations, object interest, and kinaesthetic changes (Thelen, 1981).

There is some evidence of the relationship between abnormal behaviour and welfare in other animals. In mice, early weaning is associated with both higher cortisol levels 48 hours after weaning and higher adult stereotypy levels (Würbelf and Stauffacher (1997).

In mink, pacing behaviours are documented at increasing frequency during pre-feeding which is possibly motivated by hunger (Mason, 1993). In pigs, stereotypies occur more frequently when tethered (Cronin et al., 1986; Barnett, 1990) and when housed in crates rather than in pens (Arellano, Pijoan, Jacobson, and Algers, 1992). Again, these occurred predominantly before feeding. Other behaviours such as 'vacuum' dust bathing in hens occur due to the lack of available stimuli (Lindberg, 1997)

INTERPERSONAL COMPARISONS OF UTILITY

Strength of Indicator: Low

Strength of Evidence: Weak

Abnormal behaviours occur due to abnormal physiology and therefore may not occur in all individuals exposed to the same conditions. The variation in behaviour between individuals could add considerable between-individual noise (Garner, 2005), making it very difficult to compare the degree of abnormal behaviour and then infer welfare. Mason (2004) states that 'non-stereotyping or low-stereotyping should not be overlooked or assumed to be faring well: simple measures of frequency should not be used to compare stereotypies that differ in age, form or the biological or experiential characteristics of the performing animal'.

VOCALISATION

OVERVIEW OF THE INDICATOR

Vocalisations are sounds produced by an animal. They are used to communicate between individuals. Like any behaviour, particular states or moods can cause different vocalisations. These are both an indicator of the internal state of the animal and they may also modulate emotions of those hearing the sound for example distress utterances in an abattoir (Manteuffel, Puppe, Schön, 2004). These types of calls or sounds associated with each state are built up over time by experts familiar with the animals.

Vocalisations can either be recorded by an external reviewer or more reliably and widely through automated bioacoustics. This allows for more diverse numerical descriptions and statistical examinations (Manteuffel, Puppe, Schön, 2004). The most widely known application of this sort of audio analysis is for speech recognition in smartphones (Mcloughlin, 2019).

RELIABILITY	Strength of Indicator: High
	Strength of Evidence: Conjecture
	There has been little research into the reliability of vocalisations or different automatic or reviewer based assessments of what state they indicate. Given the reliability of reviewers for other indicators and the success of audio analysis in other contexts I would expect them to be highly reliable measures of frequency and a moderately reliable measure of the type of vocalisation. We can make some weak inferences about how reliably animals produce vocalisations from other existing research. A study from Watts et al. (1999) studied the 'effects of restraint and branding on rates and acoustic parameters of vocalisation in beef cattle'. During their investigation, they found that 65 of the 189 calves studied vocalised during treatment. Given the vast majority of animals did not vocalise, this is some weak evidence that vocalisation is an unreliable indicator, at least between individuals. It's not clear whether repeated treatment would elicit the same response from the animals or whether similar groups would respond similarly.
FACE VALIDITY	Strength of Indicator: High
	Strength of Evidence: Conjecture

	Theoretically, vocalisations are a way for animals to communicate their inner mental states to other members of their own species. This might be to alert others to danger or for bonding purposes. In humans, subjective reports of well-being can to some extent be categorised as vocalisations. The complexity of the information delivered is much higher but simple vocalisations seem to provide us with a fair indicator of welfare. The main issues will be accurately mapping vocalisations onto the correct emotions and that some species or individuals will be unable to vocalise or will do so much less frequently.
CONTENT VALIDITY	Strength of Indicator: Moderate
	Strength of Evidence: Weak
	Vocalisations have been proposed as an indicator of both positive and negative welfare, or of generally high arousal affective states. Their link with negative welfare is evident from Watts et al. (1999) and similar studies that show that some animals will vocalise when in pain. For positive welfare, Boissy et al. (2007) describe volcalisations as the most promising convenient indicator for assessing positive experiences.
CRITERION VALIDITY	Strength of Indicator: High
	Strength of Evidence: Medium
	In humans, we associate many different vocalisations with different mental states. Beyond the complex information provided by speech, extreme fear or excitement can cause us to scream, humour during social interactions causes us to laugh (Bryant, 2014), and a common response to pain is to swear (Stephens, Atkins and Kingston, 2009). Vocalisations have been associated with similar experiences in non-human animals.

Similar vocalisations to laughter have been observed in rats during play and while being tickled (Panksepp and Burgdorf, 2003). The same frequency call has been associated with play, mating, and in anticipation of rewards, while lower frequency calls are exhibited during social defeat and drug withdrawal, as well as in anticipation of aversive events (Burgdorf et al., 2005).

Ewes bleat when separated from their lambs and when they are searching for each other, producing a 'rumble' when reunited (Keeling, 2001).

Cows in commercial slaughter plants vocalise after stressful events such as electric prodding, which reduces when prod use is decreased (Grandin 1998). Calves also respond to branding by vocalising at a higher peak frequency and sound level than their other vocalisations (Watts, 1999). They respond similarly, but to a much lesser degree, when separated from their mothers. There is an additional increase in vocalising and sound frequency when fed less milk (Thomas et al. 2001).

INTERPERSONAL COMPARISONS OF UTILITY

Strength of Indicator: Low

Strength of Evidence: Conjecture

As the purpose of vocalisations is to communicate between members of the same species, we would expect similar vocalisations to reflect the same emotions across individuals of the same species. However, the propensity of each individual to vocalise given the same emotional state may vary. This makes vocalisations an unreliable method to compare between two individuals. If used on a larger sample as the studies above have used, then these differences will somewhat cancel out, making it a better indicator.

However, this does not apply between species as although almost all birds and mammals are able to vocalise, the number of incidences and frequency of vocalisation may vary considerably between species (Manteuffel, Puppe, Schön, 2004). In addition, in a large number of cases species are unable to vocalise, though this is largely the case in species of fish.

BODY LANGUAGE
(QUALITATIVE BEHAVIOURAL ASSESSMENT)

OVERVIEW OF THE INDICATOR

Body language is a type of non-verbal communication through behaviours such as posture, expressions, eye contact, and gestures. For example, a dog experiencing fear or anxiety may lean back, lower its head and body, and tremble (7 Tips on Canine Body Language, 2017). What body language is being displayed and what emotion this implies is assessed by expert observers. Through their previous experience with the animals and/or training, they develop a sense for what is or is not normal behaviour and what behaviours are associated with different emotions. The scientifically validated method for assessing this is termed qualitative behavioural assessment (Qualitative Behaviour Assessment, no date). This method uses the ability of human observers to evaluate details of behaviour, posture, and the environment to code them with a descriptor such as “curious” or “happy”.

RELIABILITY	Strength of Indicator: High
	Strength of Evidence: Strong
	The inter-rater reliability of qualitative behavioural assessments has been studied in many different farmed animals. Wemelsfelder et al. (2001) investigated the inter and intra-observer reliability of qualitative assessment of pig behaviour by nine ‘naive’ observers using a free choice profile. They found significant variation across observers’ agreement showing very high inter- and intraobserver reliability. Qualitative behaviour assessments have also been shown to be reliable across different groups where one might expect bias. Wemelsfeder er al. (2012) found that pig farmers, large animal veterinarians, and animal protectionists provided consistent answers when describing the welfare of pigs (pearson $r > 0.90$).

In sheep, Phythian et al.'s (2012) observer codings for the presence of dull demeanour have high inter and intra-observer reliability. Phythian et al. (2013) also found high levels of agreement between a group of veterinary students, veterinary surgeons, and farm assurance inspectors when using a visual Analogue Scale.

In a comparison between experienced and inexperienced observers of dairy cows, Bokkers et al. (2012) found good to fair intra and inter-observer reliability. Although the author notes that some sources suggest that a correlation coefficient above 0.7 is referred to as the threshold for an acceptable measurement and this was not reached.

It is worth noting that there has been little validation of this method outside of mammals. Given the difficulty humans have empathising with other animals we are skeptical of the generalisability of these results to other animals such as fish and invertebrates.

FACE VALIDITY

Strength of Indicator: Moderate

Strength of Evidence: Conjecture

The face validity of body language varies from species to species. Generally, the more closely related the animal is to humans, the more we would expect human raters to be able to interpret body language.

CONTENT VALIDITY

Strength of Indicator: High

Strength of Evidence: Weak

The descriptors used in qualitative behavioural assessment capture a wide range of possible emotions. These include 'content/ relaxed/ bright', 'distressed/ dejected/ tense', 'agitated/ responsive/ anxious', and 'dull/ dejected/ relaxed' (Phythian et al., 2013). In some instances, a free choice model is used where observers can assign any descriptor to the animal's behaviour. This allows qualitative behavioural assessment to capture both positive and negative affect.

CRITERION VALIDITY	Strength of Indicator: Moderate
	Strength of Evidence: Medium

A forward ear posture has been associated with pain and tail docking in lambs (Guesgen *et al*., 2016). The same posture has been observed in sheep separated from its group whilst passive ear postures are more associated with enriched feed compared to wooden pellets during feeding (Reefmann *et al*., 2009; Hemsworth et al., 2011) reported a correlation between head position and serum cortisol concentration when sheep were approached by a stockperson.

Other methods such as morphometric geometrics have been associated with stereotypic or abnormal behaviour and, to a lesser degree, with depressed-like postures in horses (Sénèque et al., 2019). 'Withdrawn' posture has also been associated with greater indifference to stimuli and more emotional reactivity to challenging situations, although the animals also have lower plasma cortisol levels (Fureix et al., 2012).

Qualitative behavioural assessment has been validated against some psychological indicators demonstrating the expected associations. Examples include heart rate, heart rate variability, core body temperature, and a stress leukogram in sheep and cattle (Wickham *et al*., 2012; Stockman et al., 2011). They have also been validated in the administration of a neuroleptic drug (Rutherford et al., 2012), although Carroll et al. (2018) note that this probably results in more 'conspicuous' behaviours than usual. In their own examination of pigs, Carroll et al. (2018) found no association between qualitative behavioural assessment and psychological indicators.

INTERPERSONAL COMPARISON OF UTILITY	Strength of Indicator: High
	Strength of Evidence: Conjecture

The descriptors used during qualitative behavioural assessment are generalisable across individuals and species. The behaviours and posture of the animal that may lead an observer to ascribe these emotions onto the individual are not. These vary across species which is why this method requires previous experience and/or training in interpreting the body language of an animal.

LOCOMOTION

OVERVIEW OF THE INDICATOR

Locomotion includes any movements made by the animal to move between locations. This extends beyond abnormal behaviours such as moving and pacing around an enclosure to general movement including escape behaviour, the ability for the animal to move freely, and the frequency of these movements.

RELIABILITY	Strength of Indicator: High	
	Strength of Evidence: Weak	
	Little work has been done assessing the reliability of locomotion measures, such as distance traveled or swim speeds as a whole. However, we expect that the methods usually used to assess these are reliable. These include the stereocinematographic method which Boisclair (1992) found can accurately assess fish swim speed, or frame by frame analysis using video tracking systems (Robin Technologies, Inc. http://www. robintek.com, no date)	
FACE VALIDITY	Strength of Indicator: Low	
	Strength of Evidence: Conjecture	
	Deviations from normal movement patterns seem like a weak indicator of welfare. We may observe a reduction in locomotion due to environmental restrictions but this does not necessarily demonstrate a reduction in welfare.	
CONTENT VALIDITY	Strength of Indicator: Low	
	Strength of Evidence: Conjecture	
	Increases and decreases in the amount of movement from a normal level are both associated with negative welfare. As such, the amount of locomotion seems unable to capture positive welfare. Other movements such as escape and avoidance behaviour are also negative indicators but are already captured by preferences.	
CRITERION VALIDITY	Strength of Indicator: Low	
	Strength of Evidence: Weak	

Changes in locomotion in chickens occur due to higher stocking densities with increased density resulting in a short average distance per hour (Lewis and Hurnik, 1990). Similarly, Fouad et al. (2008) observed that floor-raised broiler chickens were seen more often walking, lying, and pecking compared to cage-raised chickens who stood still more often. Shields and Greger (2013) notes that 'the lack of free space appears to constrain activities that broiler chickens would otherwise choose'. The cage-raised chickens had greater heterophil to lymphocyte ratios and worse gait scores.

Kristiansen et al. (2004) found that in Atlantic halibut individual swimming activity rose with increasing density. Fish that were frequent "surface swimmers" also had significantly lower growth but this may be classified as a stereotypic behaviour. Changes in swimming activity have also been associated with hypoxic conditions for different species of fish. This can lead to either reduced activity which 'may enable fish to survive prolonged and widespread exposure'(Martins et al., 2011) or increased swim speed which may be an escape response (Tang and Boisclair, 2011).

INTERPERSONAL COMPARISON OF UTILITY

Strength of Indicator: Low

Strength of Evidence: Conjecture

Comparisons between individuals are difficult as swim speed, average distance travelled, and other movement patterns will vary naturally based on each individual's temperament rather than their well-being. Some individuals will be more prone to activity and more effective when it is constrained. Therefore comparing the change from standard movement patterns can't provide us with an adequate comparison. Species comparisons are impossible for some movement patterns such as flight or swim speed and others will vary greatly between animals that would naturally roam large territories.

NATURAL BEHAVIOURS

OVERVIEW OF THE INDICATOR

As we discussed in the introductory section on ‘what is animal welfare’ many view the ability to perform natural behaviours as essential to the welfare of the animal. So much so, that they argue that the concept of welfare itself contains the natural behaviours of the animal. Although we think this is philosophically flawed, the performance of a wide variety of natural behaviours are a potential indicator of good welfare.

The definition of what is and is not a natural behaviour is itself contentious. Bracke and Hopster (2006) examine previous perspectives on what is or is not classified as natural behaviour and their flaws. They conclude that the proper definition of a natural behaviour is a ‘compound working-definition’ and that ‘Natural behaviour is behaviour that animals tend to perform under natural conditions, because it is pleasurable and promotes biological functioning’.

Behaviours classified as natural or comfort behaviours differ between species. Under this definition, examples of natural behaviour include rooting and nest building in pigs, dust-bathing and preening in poultry, grazing in cattle, and play behaviour in all animals (Bracke and Hopster, 2006).

RELIABILITY	Strength of Indicator: High
	Strength of Evidence: Weak
	The most common methods for assessing the frequency of behaviours is the use of an ethogram by human observers. The reliability of this kind of coding in general was examined in several previous sections of this report. Other methods for coding behaviour include using machine learning and at least in some cases these methods are reliable (Pereira et al., 2013). However, in this review there was little direct research into the reliability of ethograms for coding natural behaviours.
FACE VALIDITY	Strength of Indicator: High
	Strength of Evidence: Conjecture

	Natural behaviours have high face validity as an indicator of positive welfare as they demonstrate time spent experiencing positive affect. The frequency of these behaviours show that the animal is experiencing at least some moments of happiness and that they have high enough welfare to be motivated to perform these. The widespread belief that natural behaviour is a constituent of welfare also provides some evidence in favour of its validity as an indicator. However, for some definitions of natural behaviour this may be because the public overvalues it due to naturalistic bias.
CONTENT VALIDITY	Strength of Indicator: Low
	Strength of Evidence: Conjecture
	Natural behaviour is an indicator of positive welfare experiences. Time spent performing these activities as well as their diversity are associated with positive affect. We can evaluate the intensity of these experiences by examining the animal's response to the frustration of their desire. The degree of frustration will be associated with the intensity of the desire which will somewhat map onto the level of reward provided by the behaviour.
CRITERION VALIDITY	Strength of Indicator: High
	Strength of Evidence: Weak
	Various natural and comfort behaviours have been associated with physiological responses linked with positive emotions. Play, autogrooming, and dust-bathing are all demonstrated signs of rewarding activity (Vestergaard et al., 1999; Spruijt, van Hooff and Gispen, 1992). Individuals show compensatory behaviour when behaviours such as dust-bathing are restricted (Vestergaard et al., 1999). This demonstrates that animals find these behaviours rewarding and are frustrated when they are unable to perform them, but provides no indication to whether high welfare individuals are more likely to perform them.

In this brief review, we found several studies which provide evidence that stressors or environments associated with lower welfare reduce the incidence of natural behaviour.

Mintline et al. (2013) found that play behaviour in calves was suppressed 3 hours after hot-iron disbudding. Administering local anaesthetic and a non-steroidal anti-inflammatory drug increased play behaviour in the disbudded group. This treatment had no effect on the play behaviour of calves in the sham disbudding group.

Pohle and Cheng (2009) examined the different effects of furnished cages and battery cages on bird behaviours. They found that the birds housed in furnished cages had higher levels of preening. The birds housed in battery cages had higher posture and behavioural transitions and increased time spent walking and performing exploratory behaviour which Pohle suggests may indicate they were stressed.

Bergmann et al. (2017) compared the effect of an alternative rearing system for broiler chickens against conventional methods. The chickens raised in the alternative system had access to several forms of environmental enrichment and a lower stocking density. They found that the birds raised in this system made good use of the enrichment display, spent more time grooming but that there was no significant increase in dust-bathing. By contrast Liu et al. (2020) found that chicks in enriched enclosures were less likely to display play behaviour during worm running tests, though they found no difference in spontaneous play behaviour between groups. Liu suggests that the increase in worm run play behaviour in the unenriched group may be ‘because of the larger contrast between their relatively unstimulating environment and the test’ (Liu, Torrey, Newberry, & Widowski, 2020).

Miller et al. (2020) suggests that the diversity of behaviours should also be used as a positive indicator of welfare. Through their review they outline evidence that behavioural diversity is negatively associated with stereotypies, fecal glucocorticoid metabolites, and poor health while it increases with environmental enrichment.

INTERPERSONAL COMPARISONS OF UTILITY

Strength of Indicator: Low

Strength of Evidence: Conjecture

As was discussed above, there are a wide variety of natural behaviours and some behaviours such as preening are only seen in a few species. When we have established the natural behaviour patterns and the level of motivation the individual has to perform each behaviour (Bracke and Hopster, 2006), we are able to compare the degree of frustration. However, the natural variation in personality between individuals will affect the desire to and propensity to perform natural behaviours in different systems, making inter-individual comparisons difficult. As with many of the indicators discussed, the stated solution is to use a large enough sample such that these personality differences even out.

OVERALL ASSESSMENTS OF ANIMAL WELFARE

Image Credit:Fran Hogan

Although indicators are useful tools for assessing welfare, most are not in and of themselves an overall assessment of the animal's welfare. Preference testing, cognitive bias, and qualitative behavioural assessment can be used as standalone assessments but primary biological indicators or vocalisations only reveal part of the picture. There are a large variety of existing composite measures of welfare used by non-profits and used in academia. These vary both in the weight they put on different indicators and in the underlying philosophy of what animal welfare is, as discussed in the first section of this report. The ideal overall assessment of welfare would be a ratio scale where there is a true zero point that reflects a neutral life and equal intervals between points.

FIVE FREEDOMS

The five freedoms is one of the most widely used methods for assessing welfare. They were developed by Britain's Farm Animal Welfare Council in 1965 for aspects needed to meet the mental and physical needs of animals (Miller et al., 2020). The five freedoms constitute the following: freedom from hunger and thirst, freedom from discomfort, freedom from pain injury or disease, freedom to express normal behaviour, and freedom from fear and distress. These are assessed in part through different indicators of welfare and via an objective list of what constitutes high welfare for an animal. Freedom from thirst, for example, is assessed through access to fresh water rather than psychological measures of dehydration.

Existing concerns with the five freedoms critique its focus on poor welfare, which may be sufficient to reduce the suffering of animals but may not result in a life worth living. The only freedom that touches on positive welfare is the freedom to express normal behaviour. Even in this, it is only the freedom to express these behaviours rather than the extent to which the animal does. The five freedoms model would fail to capture the value of the types of behaviours expressed or the variety, which may also be an indicator of positive welfare (Miller et al., 2020). Overall, the five freedoms are more a guide for how to avoid negative welfare rather than a measure of the overall quality of an animal's life and a method to achieve flourishing. Each freedom is necessary for good welfare but not sufficient. In addition, the five freedoms don't provide any way to grade the severity

of the frustration of a freedom (McCulloch, 2012).

THE FIVE DOMAINS MODEL

The five domains is a more recent iteration of the five freedoms model for welfare. This method reformats the five freedoms into domains of nutrition, environment, health, behaviour and mental state. These are graded on a five tier system for welfare from A-E, where A represents the highest welfare and E the lowest, alongside a grading of animal welfare enhancement using a four-tier scale (0, +, ++, +++) (Mellor, 2017). These grades represent 'different degrees of welfare compromise ranging from none to very severe'(Mellor, 2017). Grades are intended as an ordinal scale with no intention of weighing each domain's importance against each other. Mellor (2017) explains that the rejection of numerical grading was to avoid non-reflective averaging of "scores" and to avoid implying much greater precision than is possible with qualitative assessments. This means that the five domains make no comparative claims between two animals, one with grade A on nutrition and B on environment, and the other with grade A on environment and B on nutrition. Therefore, as it is originally construed it can't be used to prioritize between interventions on its own, and instead it is meant to be used to help inform one's considered judgment.

FIVE PROVISIONS MODEL

The five provisions model is a further iteration of the five freedoms. The provisions acknowledge that complete freedom from all negative experiences is an unrealistic ideal; Mellor (2016) argues that the best that can be achieved is to minimise them. The provisions also hope to improve on the common critique that the five freedoms focus too much on negative experiences. It does this by including a provision for positive mental experiences and placing a strong emphasis on the subjective experience of the animal, with the aim of achieving a net balance between significant negative and positive experiences.

TWELVE CRITERIA (KEELING, 2007)

Botreau's twelve criteria is made up of four criteria (good feed, good housing, good health, and appropriate behaviour) and 12 sub-criteria. These criteria were based on an additional critique of the five freedoms provided by Botreau that many of the freedoms overlap or are very vague (Keeling, 2007). For example, freedom from discomfort is often associated with freedom from pain, injury, and disease.

QUALITY OF LIFE (MCMILAN, 2003)

McMillan (2003) grounds their assessment of welfare in the balance model of quality of life. This is the affect balance between pleasant and unpleasant feelings at each point in time and across an animal's life. The quality of one's life is assessed by each individual based on each animal's genetics, personality, and experiences, which results in different values and priorities being assigned by the individual to different aspects of its life (McMillan, 2003). McMillan outlines six major contributing factors to an individual's quality of life: social relationships, mental stimulation, health, food consumption, stress, and

control. These are outlined with methods for maximisation but no overall assessment or prioritisation between different elements is likely given the individual variation in weightings.

QUALITY OF LIFE DOMAINS (TAYLOR AND MILLS, 2007)

Taylor and Mills (2007) and Teng et al. (2018) define quality of life as the state of an individual's life as perceived by them at one point in time. This includes the balance of positive and negative affect and cognitive evaluations when the animal has the capacity. They acknowledge that quality of life can be predicted by the fulfillment of various needs but that these are not perfect proxies. They outline various domains that can be used as indicators of an animal's quality of life. The two broad categories used are social or environmental indicators and physical or psychological indicators. These were developed from child–proxy health related quality of life tools and existing companion animal quality of life assessments.

WELFARE-ADJUSTED LIFE YEARS (WALYs)

Welfare-Adjusted Life Years (WALY) are an adaptation of the widely used global health metrics disability adjusted life years (DALY). The WALYs lost due to an event or condition are calculated from the welfare weight of the condition or event multiplied across the period of time that the animal is affected and the years of life lost. The welfare weight is constructed by surveying veterinarians or animal welfare experts for a particular animal on the welfare weight they would assign to various conditions. This can either be done using a visual analog scale or through time trade-offs between different conditions and years of life. Teng et al. (2018) introduced this metric by estimating the WALYs for 10 canine diseases.

SEMANTIC MODELLING OF EXPERT OPINION

Bracke et al. (2019) used an alternative method for synthesising expert opinion into a quantitative score for different housing systems for broiler chickens. Bracke surveyed 'established welfare scientists and others (e.g. veterinarians) to provide welfare scores and weighting factors for welfare-relevant attributes/parameters of broiler housing systems' (Bracke et al., 2019). This included overall scores, welfare scores for different housing systems, and the main parameters of these systems with their relative importance.

CHARITY ENTREPRENEURSHIP'S WELFARE INDEX

Charity Entrepreneurship's Welfare Index is a weighted index of 8 criteria used to assess welfare (Is it better to be a wild rat or a factory farmed cow? A systematic method for comparing animal welfare, no date). These are death rate/reason, human preference from behind the veil of ignorance, disease/injury/functional impairment, thirst/hunger/malnutrition, anxiety/fear/pain/distress, environmental challenge, index of biological markers and behavioural/interactive restriction. These are each weighted according to their performance on CE's evaluation criteria for the underlying goals of the metrics. Each area is assessed for the animal and then

given a score in the range assigned to each criteria by multiple researchers. This is informed by the range of the criteria in different environments or for different conditions.

ANIMAL NEEDS INDEX (ANI)

The Austrian ANI-system (ANI-35-L) evaluates five components: mobility, social contact with members of the same species, floor conditions, stable climate, and the intensity of human care ('A review of the animal needs index (ANI) for the assessment of animals' well-being in the housing systems for Austrian proprietary products and legislation', 1999). Within these components, there are 24 criteria graded up to a maximum of plus 3 and a minimum of -0.5. These criteria are all aspects of the environment or about access to goods rather than animal-based measures. This means that ANI needs to be adapted for use for each new species. This has already been done in several species such as beef cattle, dairy cows (Seo, Date, Daigo, Kashiwamura and Sato, 2007), and laying hens but is not available for others. Once adapted for a new animal, the ANI provides a quick way to assess the conditions for welfare but the lack of animal-based measures makes it a poor guide for the importance of different aspects of an animal's environment or care.

WELFARE QUALITY PROJECT

The Welfare Quality project was a European research project which aimed to develop assessment protocols for animal welfare (Blokhuis et al., 2010). Welfare Quality outlines four principles and twelve welfare criteria for assessing welfare. Each criteria is measured through a variety of indicators (~30) which vary between species, with different protocols available for each (Welfare Quality Network, no date). Many of these indicators are animal-based measures, such as physical health or behaviour, but the WQI also includes items such as the number of drinkers available per animal (Quality®, no date). Compliance with different measures of each criterion is expressed on a 0-100 scale; these are amalgamated into criteria scores which are then amalgamated into principle scores using a choquet integral. An intuitive way to understand this function is it weights the lowest score more highly, so if one welfare criteria or attribute is compromised the overall score can't be compensated for by good scores in other sections.

ANIMAL WELFARE INDICATORS PROJECT

Animal Welfare Indicators (AWIN) is another research project from the European Union (European Commission, 2015). AWIN is an evolution of the WQI with a greater focus on animal-based measures and assessing pain, as this was highlighted as an area neglected by existing measures. Resource or management based indicators are only used where no valid, reliable and feasible animal-based measures exist. In addition, the AWIN project focused on animals not covered by the WQI, such as horses and turkeys.

SOWEL AND SOWEL-TYPE MODELS

The SOWEL (from SOw WELfare) models were originally developed for the measurement of pregnant sows by Bracke et al. (2002) In sows, the model measures a list

of 11 welfare needs with each need covering one of the animal's behavioural systems such as feed intake, thermoregulation, rest, and locomotion. Each need is made up of different attributes,
which are a mixture of environment or animal-based or management related measures, which are scored between 0 and 1 according to pre-set criteria, where 1 represents the best score and 0 the worst. Intermediate scores are equally distributed and assigned based on the number of levels for that attribute so that an attribute with three levels receives attribute scores of 0, 0.5, and 1.

These attribute scores are then weighted according to a list of 12 weighting categories which classify welfare performance criteria from various welfare disciplines. These categories are a selection of different animal-based welfare indicators capturing all the dimensions examined in previous sections: preference-based, physical health, psychological and behavioural. The weighting for each attribute level of each attribute is calculated from the sum of the scores from each category. This is informed by numerous scientific statements from different effects of the attribute on different welfare indicators to observations on natural behaviour. This results in a different weighting depending on the level of each attribute such that the difference between a large amount of space per pen (> 6,250 m2) and a moderate amount may be lower than a small amount (1–1.5 m2) and a moderate amount. The score for each housing condition is the product of the attribute scores and the weightings.

MYFISHCHECK

The most recently developed overall assessment is MyFishCheck which is an evolution of SOWEL type semantic models. This is a model of welfare made up of natural living, physical functioning, and a feelings based aspect. These are measured through five modules: farm management, water quality, fish group behaviour, fish external appearance, and fish internal appearance. Within each module, there are multiple parameters each with parameter intervals set through a literature review. Each interval is given a parameter score between 0 and -1 equally distributed across intervals as with SOWEL type models. These parameter scores are then multiplied by the score weights, which range from 1 to 5 depending on the frequency and severity of the stressor. These again are then multiplied by their parameter weights from 1 to 5, defined by interviewing 20 experts on the relevance of the parameter for the welfare of the animal. The overall score is calculated through the weighted average of all parameter scores, weighted by score and parameter weights and adding one.

DISCUSSION

The problems with many of these systems are that they are implicitly grounding their theory of animal welfare in the objective list theory of well-being. This theory, as we discussed in the first section of this report, has numerous flaws as a theory of animal welfare. In the context of many of these systems, this manifests when they, for example, list housing conditions as a condition for good welfare. This will be linked to welfare in many cases, but to have confidence in this claim the welfare status of each housing system needs to be validated against available indicators rather than asserted as a principle. Other authors such as McMillan (2003) explicitly state that their underlying theory of well-being is hedonism-based or grounded in the subjective experience of the animal. The criteria they use are proxies for this underlying well-being rather than criteria for welfare. This approach seems much more appropriate given our theory of well-being is grounded in hedonism and desire theories of well-being.

For these systems, thought needs to be put into ensuring the proxies used are reliable and valid indicators of well-being or welfare and that they are weighted based on their strength. Some explicitly place weight on each of these criteria but the vast majority provide no weights and therefore de facto weigh all factors equally. Many of these systems also fail to provide rubrics or other systems for evaluating how well an individual is doing on any given criteria. The rationale given behind this is that providing scoring rubrics or numerical scores results in undue confidence in the assessment. Therefore, at best, grading systems should be used. However, in doing so this leaves the overall impression of welfare fully up to the reviewer's judgment, rather than grounding it more in the best guess of each aspect's importance based on existing evidence.

There are multiple systems that attempt to weigh different welfare problems and indicators including Charity Entrepreneurship's Welfare Index (Is it better to be a wild rat or a factory farmed cow? A systematic method for comparing animal welfare, no date), WALYs (Teng et al., 2018), ANI, the Welfare Quality Project and SOWEL-type models. The Welfare Index is itself an evolution of other methods like the five domains with an additional attempt to weigh different criteria. The main weakness with this system are the concerns about the validity of the relative weight placed on each criteria, the use of purely negative indicators like mortality as positive indicators, and the reliance on the reviewer's judgment for the score given to each criteria. Additionally, the criteria used fall somewhat into the objective list like environmental challenge, while others are purely used as indicators of welfare such as the index of biological markers of happiness.

Some of these concerns can be mitigated by reweighting the criteria (potentially for each species) and adjusting the range for different indicators. Unfortunately, the reliance on the reviewer's judgement for the score for each criteria cannot currently be avoided. If we had significantly more data, we could model a range of indicators in the same group of animals and use factor analysis to establish how much each indicator loads onto what we would assume

is well-being. We could use this to help assign the weight we give to each indicator in the index. With additional data, we could also assess the inter-rater reliability of what is ostensibly a visual analog scale for the performance on each indicator.

Like the Welfare Index, Welfare-Adjusted Life Years (WALYs) and Bracke's Expert Survey both rely on individuals' judgment of visual analog or likert scales. However, in this case judgments are provided by Veterinary and animal welfare specialists rather than generalist researchers. Still, there has been no research into the inter-rater reliability of these assessments other than to note that in Teng et al. (2018) the use of time trade offs (TTO) or a visual analogy scale resulted in similar welfare weights. The additional problems with this method are that similar methods used in humans don't perfectly track the subjective reports of those with the condition (A Happiness Manifesto: Why and How Effective Altruism Should Rethink its Approach to Maximising Human Welfare - EA Forum, no date). This problem will likely be exacerbated for other animals, particularly those less closely related to us as their subjective experiences and preferences can be dramatically different than our own. Given these problems, these two systems are better viewed as a qualitative representation of expert opinion.

The Animal Needs Index is one of the weaker models examined for the measurement of welfare. Once established and validated, it is a quick easy proxy for farmers or inspectors but its reliance on environmental measures means it can't be used to easily compare new aspects and potential improvements to an animal's environment, such as advancements in environmental enrichment.

The Welfare Quality projects shares similar features with the Animal Needs Index in that many of its measures are environmental. The main source for the overall welfare score comes from the measure scores used to evaluate different criteria. These scores are calculated a variety of ways, including decision trees, weighted sums based on the severity and frequency of different conditions, and alarm thresholds. As Browning (2019b) has commented, the reasoning behind 'aggregation weightings are also quite opaque, and seem to be based on expert opinion rather than measured effect on the animals' . If the reasoning behind these weightings were more transparent, and could be more easily updated based on future research, this would be a stronger system. As with the ANI, this makes it harder to work with when investigating new advancements in animal care. One potential strength of this system is the use of a Choquet integral which seems to better capture the intuition that a failure in one aspect of an animal's environment or health can't be compensated by good performance on other criteria.

SOWEL-type models are a marked improvement on many other systems in that they are transparent in their weightings of different criteria and method of scoring. For this reason, I expect that inter-rater reliability would be high as the system is doing more of the work in providing a score. Browning (2019a) describes one weakness of this system in that the attribute level scores are arbitrary. The model implicitly assumes that the difference

between attribute levels for the welfare score are equal, but for many attributes this is likely not the case. Another concern raised by Browning with this system is the range of information used for weighting the importance of different attributes. This includes anything from hard data on the effect of reliable indicators to expert opinion. This can result in a wide range of confidence in the weights assigned to different attributes. However, this can also be viewed as a strength of the system in that we are able to update our views and thus the score of different housing systems based on a wide variety of possible evidence. As long as the system remains transparent about relative confidence in different attributes, this isn't a great concern.

An additional weakness of SOWEL-type models is with their weighting systems. Although, as Browning has highlighted the welfare score is a potential source of arbitrary scoring, the weightings are as well. The use of maximum and minimum values in the weighting of different attributes levels leaves no room for new research using an existing type of weighting category to update the system. For example, if there are two operant test studies showing chickens' willingness to pay for cage free over caged systems, one with a high value and the other a low one. If a new study replicating this demonstrated a low willingness to pay marginally above the lowest study but not significantly, this wouldn't update the system towards a lower welfare estimate. MyFishCheck suffers from the weaknesses of other SOWEL-type semantic models in that parameter scores are arbitrarily set as equidistant within the range. For this reason, it is not an improvement on SOWEL-type models. The main benefit is that it is more operationable, even if this comes at the expense of accuracy.

In conclusion, all of the welfare indicators and overall assessments of welfare reviewed in this report have numerous flaws. Establishing a valid objective assessment for the welfare of individuals is a very difficult task and no single system provides the “silver bullet” for measuring welfare. The welfare assessments of every system need to be held in the context of each system's flaws and methods of calculation, and interpreted with this in mind. Instead, a better method for assessing welfare would be a composite of multiple qualitative and quantitative methods which each provide a different perspective on welfare. If multiple systems and methods of assessments converge on which welfare improvements we should prioritise, we can have greater confidence than relying on any given measure. The ideal combination of measures would be some combination of a qualitative measure, expert opinion based measures, an index or semantic model of animal-based measures and standalone measures, such as preference testing or qualitative behavioural assessment. This could be the combined use of the Five Domains, a semantic model of expert opinion, a SOWEL-type model and any standalone measures, such as preference testing or qualitative behavioural assessment. In practice, the cost of using such an extensive list will make it impractical for many decision-makers, but this could be the basis for large-scale asks. A more limited combination would be the use of the Five Domains, Charity Entrepreneurship Welfare Index, any standalone measures and interviews with experts.

BIBLIOGRAPHY

7 Tips on Canine Body Language (2017). Available at: https://www.aspcapro.org/resource/7-tips-canine-body-language (Accessed: 15 March 2021).

2017 Report on Consciousness and Moral Patienthood (2018). Available at: https://www.openphilanthropy.org/2017-report-consciousness-and-moral-patienthood (Accessed: 15 March 2021).

Abdallah, T. (1998) 'The Satisfaction with Life Scale (SWLS): Psychometric Properties in an Arabic-speaking Sample', International journal of adolescence and youth, 7(2), pp. 113–119. doi: 10.1080/02673843.1998.9747816.

A. C. Lindberg, C. J. N. (1997) 'Dustbathing in modified battery cages: Is sham dustbathing an adequate substitute?', *Applied animal behaviour science*, 55(1-2), pp. 113–128. doi: 10.1016/S0168-1591(97)00030-0.

'Age and weight at weaning affect corticosterone level and development of stereotypies in ICR-mice' (1997) *Animal behaviour*, 53(5), pp. 891–900. doi: 10.1006/anbe.1996.0424.

A Happiness Manifesto: Why and How Effective Altruism Should Rethink its Approach to Maximising Human Welfare - EA Forum (no date). Available at: https://forum.effectivealtruism.org/posts/FvbTKrEQWXwN5A6Tb/a-happiness-manifesto-why-and-how-effective-altruism-should (Accessed: 12 March 2021).

Aishvarya, S. *et al*. (2014) 'Psychometric properties and validation of the Satisfaction With Life Scale in psychiatric and medical outpatients in Malaysi', *Comprehensive psychiatry*, 55(1), pp. 101–S106. doi: 10.1016/j.comppsych.2013.03.010.

Akre, A. K., Bakken, M. and Hovland, A. L. (2009) 'Social preferences in farmed silver fox females (Vulpes vulpes): Does it change with age?', *Applied animal behaviour science*, 120(3), pp. 186–191. doi: 10.1016/j.applanim.2009.06.008.

Amdam, G. V. (2011) *Measuring Animal Preferences and Choice Behavior*, *Nature Education Knowledge*. Available at: https://www.nature.com/scitable/knowledge/library/measuring-animal-preferences-and-choice-behavior-23590718/ (Accessed: 12 March 2021).

Ana Xavier, M. C. A. J. P. G. (2014) 'Deliberate self-harm in adolescence: The impact of childhood experiences, negative affect and fears of compassion', *Revista de Psicopatología y Psicología Clínica*, 20, pp. 41–49. doi: 10.5944/rppc.vol.1.num.1.2015.14407.

A Piller, S Bergmann, A Schwarzer, M Erhard, J Stracke, B Spindler, N Kemper,P Schmidt, J Bachmeier, B Schade, B Boehm, E Kappeand H Louton (2020) 'Validation of histological and visual scoring systems for foot-pad dermatitis in broiler chickens', *Animal welfare*, 29, pp. 185–196. doi: 10.7120/09627286.29.2.185.

Arellano, P.E., Pijoan, C., Jacobson, L.D. and Algers, B. (1992) 'Stereotyped behaviour, social interactions and suckling pattern of pigs housed in groups or in single crates', *Applied animal behaviour science*, 35(2), pp. 157–166. doi: 10.1016/0168-1591(92)90006-W.

'A review of the animal needs index (ANI) for the assessment of animals' well-being in the housing systems for Austrian proprietary products and legislation' (1999) *Livestock Production Science*, 61(2-3), pp. 179–192. doi: 10.1016/S0301-6226(99)00067-6.

Atkinson, H. C. and Waddell, B. J. (1997) 'Circadian variation in basal plasma corticosterone and adrenocorticotropin in the rat: sexual dimorphism and changes across the estrous cycle', *Endocrinology*, 138(9), pp. 3842–3848. doi: 10.1210/endo.138.9.5395.

Barger, S. D., Donoho, C. J. and Wayment, H. A. (2009) 'The relative contributions of race/ethnicity, socioeconomic status, health, and social relationships to life satisfaction in the United States', *Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation*, 18(2), pp. 179–189. doi: 10.1007/s11136-008-9426-2.

Barton, B. A. (2002) 'Stress in fishes: a diversity of responses with particular reference to changes in circulating corticosteroids', *Integrative and comparative biology*, 42(3), pp. 517–525. doi: 10.1093/icb/42.3.517.

Bateson, M. *et al*. (2011) 'Agitated Honeybees Exhibit Pessimistic Cognitive Biases', *Current biology: CB*, 21(12), p. 1070. doi: 10.1016/j.cub.2011.05.017.

Bateson, M. (2016) 'Cumulative stress in research animals: Telomere attrition as a biomarker in a welfare context?', *BioEssays: news and reviews in molecular, cellular and developmental biology*, 38(2), p. 201. doi: 10.1002/bies.201500127.

Bergmann, S., Schwarzer, A., Wilutzky, K., Louton, H., Bachmeier, J., Schmidt, P., Erhard, M., Rauch, E (2017) 'Behavior as welfare indicator for the rearing of broilers in an enriched husbandry environment—A field study', *Journal of veterinary behavior: clinical applications and research: official journal of: Australian Veterinary Behaviour Interest Group, International Working Dog Breeding Association*, 19, pp. 90–101. doi: 10.1016/j.jveb.2017.03.003.

Berridge, K. C. (2009) 'Wanting and Liking: Observations from the Neuroscience and Psychology Laboratory', *Inquiry* , 52(4), p. 378. doi: 10.1080/00201740903087359.

Blais, M. R., Vallerand, R. J., Pelletier, L. G., & Brière, N. M. (1989) 'L'échelle de satisfaction de vie: Validation canadienne-française du "Satisfaction with Life Scale."', *Canadian Journal of Behavioural Science*, 21(2), pp. 210–223. doi: 10.1037/h0079854.

Blokhuis, H. J. *et al*. (2010) 'The Welfare Quality Project and Beyond: Safeguarding Farm Animal Well-Being', *Acta agriculturae Scandinavica. Section A, Animal science*, 60(3), p. 129–140. doi: 10.1080/09064702.2010.523480.

de Boer, E. A. M. B. M. de V. I. A. I. J. M. (2012) 'Inter- and intra-observer reliability of experienced and inexperienced observers for the Qualitative Behaviour Assessment in dairy cattle', *Animal Welfare*, 21, pp. 307–318. doi: 10.7120/09627286.21.3.307.

Boisclair, D. (1992) 'An Evaluation of the Stereocinematographic Method to Estimate Fish Swimming Speed', *Canadian Journal of Fisheries and Aquatic Sciences*, 49(3), pp. 523–531. doi: 10.1139/f92-062.

Boissy, A. *et al*. (2007) 'Assessment of positive emotions in animals to improve their welfare', *Physiology & behavior*, 92(3), pp. 375–397. doi: 10.1016/j.physbeh.2007.02.003.

von Borell, E. *et al*. (2007) 'Heart rate variability as a measure of autonomic regulation of cardiac activity for assessing stress and welfare in farm animals -- a review', *Physiology & behavior*, 92(3), pp. 293–316. doi: 10.1016/j.physbeh.2007.01.007.

Bracke, M. B. M. *et al*. (2002) 'Decision support system for overall welfare assessment in pregnant sows B: Validation by expert opinion', *Journal of animal science*, 80(7), pp. 1835–1845. doi: 10.2527/2002.8071835x.

Bracke, M. B. M. *et al*. (2019) 'Broiler welfare trade-off: A semi-quantitative welfare assessment for optimised welfare improvement based on an expert survey', *PloS one*, 14(10). doi: 10.1371/journal.pone.0222955.

Bracke, M. B. M. and Hopster, H. (2006) 'Assessing the Importance of Natural Behavior for Animal Welfare', *Journal of agricultural & environmental ethics*, 19(1), pp. 77–89. doi: 10.1007/s10806-005-4493-7.

Bradshaw, W. (2019) 'Assessing biomarkers of ageing as measures of cumulative animal welfare', *Wild Animal Iniative*. Available at: https://www.wildanimalinitiative.org/blog/biomarkers-cumulative-welfare (Accessed: 12 March 2021).

Brijs, J. *et al*. (2018) 'The final countdown: Continuous physiological welfare evaluation of farmed fish during common aquaculture practices before and during harvest', *Aquaculture* , 495, pp. 903–911. doi: 10.1016/j.aquaculture.2018.06.081.

Browning, H. (2019a) *If I Could Talk to the Animals: Measuring Subjective Animal Welfare*. Australian National University. Available at: https://philpapers.org/rec/BROIIC-3.

Browning, H. (2019b) *IF I COULD TALK TO THE ANIMALS: MEASURING SUBJECTIVE ANIMAL WELFARE*. Edited by The Australian National University. Doctor of Philosophy.

The Australian National University. Available at: https://openresearch-repository.anu.edu.au/bitstream/1885/206204/1/Browning%20Thesis%202020.pdf (Accessed: 15 March 2021).

Buchanan, T. W., al'Absi, M. and Lovallo, W. R. (1999) 'Cortisol fluctuates with increases and decreases in negative affect', *Psychoneuroendocrinology*, 24(2), pp. 227–241. doi: 10.1016/s0306-4530(98)00078-x.

Burgdorf, J. *et al*. (2005) 'Breeding for 50-kHz Positive Affective Vocalization in Rats', *Behavior genetics*, 35(1), pp. 67–72. doi: 10.1007/s10519-004-0856-5.

Carlander, D., Wilhelmson, M. and Larsson, A. (2001) 'Limited Day to Day Variation of IgY Levels in Eggs from Individual Laying Hens', *Food and agricultural immunology*, 13(2), pp. 87–92. doi: 10.1080/09540100120055657.

Carragher, J. F. and Rees, C. M. (1994) 'Primary and secondary stress responses in golden perch, Macquaria ambigua', *Comparative biochemistry and physiology. Part A, Physiology*, 107(1), pp. 49–56. doi: 10.1016/0300-9629(94)90272-0.

Carroll, G. A. *et al*. (2018) 'Identifying physiological measures of lifetime welfare status in pigs: exploring the usefulness of haptoglobin, C- reactive protein and hair cortisol sampled at the time of slaughter', *Irish veterinary journal*, 71, p. 8. doi: 10.1186/s13620-018-0118-0.

Cavigelli, S. A., Monfort, S. L., Whitney, T. K., Mechref, Y. S., Novotny, M., & McClintock, M. K. (2005) 'Frequent serial fecal corticoid measures from rats reflect circadian and ovarian corticosterone rhythms', *Journal of Endocrinology*, 184(1), pp. 153–163. doi: 10.1677/joe.1.05935 .

Chang, E. C., Asakawa, K., & Sanna, L. J. (2001) 'Cultural variations in optimistic and pessimistic bias: Do Easterners really expect the worst and Westerners really expect the best when predicting future life events?', *Journal of Personality and Social Psychology*, 81(3), pp. 476–491. doi: 10.1037/0022-3514.81.3.476.

Cheung, F. and Lucas, R. E. (2014) 'Assessing the validity of single-item life satisfaction measures: results from three large samples', *Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation*, 23(10), pp. 2809–2818. doi: 10.1007/s11136-014-0726-4.

Colborn, D. R. *et al*. (1991) 'Responses of cortisol and prolactin to sexual excitement and stress in stallions and geldings', *Journal of animal science*, 69(6), pp. 2556–2562. doi: 10.2527/1991.6962556x.

Comparing the heart rates of animals and human beings (no date) *Open Learn*. Available at: https://www.open.edu/openlearn/ocw/mod/oucontent/view.php?id=77548&section=8 (Accessed: 12 March 2021).

Contreras-Aguilar, M. D. *et al*. (2019) 'Application of a score for evaluation of pain, distress and discomfort in pigs with lameness and prolapses: correlation with saliva biomarkers and severity of the disease', *Research in veterinary science*, 126, pp. 155–163. doi: 10.1016/j.rvsc.2019.08.004.

Courtney Beard, N. A. (2009) 'Interpretation in Social Anxiety: When Meaning Precedes Ambiguity', *Cognitive therapy and research*, 33(4), p. 406. doi: 10.1007/s10608-009-9235-0.

Crisp, R. (2017) 'Well-Being', *The Stanford Encyclopedia of Philosophy*. Fall 2017. Edited by E. N. Zalta. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/fall2017/entries/well-being/.

Cronbach, L. J. and Meehl, P. E. (1955) 'Construct validity in psychological tests', *Psychological Bulletin*, pp. 281–302. doi: 10.1037/h0040957.

Cronin, G. M. *et al*. (1986) 'The influence of degree of adaptation to tether-housing by sows in relation to behaviour and energy metabolism', *Animal science* , 42(2), pp. 257–268. doi: 10.1017/S0003356100017979.

Cronin, G. M. *et al*. (1991) 'The welfare of pigs in two farrowing/lactation environments: cortisol responses of sows', *Applied animal behaviour science*, 32(2), pp. 117–127. doi: 10.1016/S0168-1591(05)80036-X.

Croyle, K. L. (2000) 'Characteristics associated with a range of self -harm behaviors in university undergraduates', . *Graduate Student Theses, Dissertations, & Professional Papers*. Available at: https://scholarworks.umt.edu/cgi/viewcontent.cgi?article=11649&context=etd (Accessed: 15 March 2021).

Dawkins, M. (1977) 'Do hens suffer in battery cages? environmental preferences and welfare', *Animal behaviour*, 25, pp. 1034–1046. doi: 10.1016/0003-3472(77)90054-9.

Decina, C. *et al*. (2019) 'Development of a Scoring System to Assess Feather Damage in Canadian Laying Hen Flocks', *Animals : an open access journal from MDPI*, 9(7). doi: 10.3390/ani9070436.

Diener, E. *et al*. (1985) 'The Satisfaction With Life Scale', *Journal of personality assessment*, 49(1), pp. 71–75. doi: 10.1207/s15327752jpa4901_13.

Dolan, P. and Metcalfe, R. (2012) 'Valuing Health: A Brief Report on Subjective Well-Being versus Preferences', *Medical Decision Making*, 32(4), pp. 578–582. Available at: https://journals.sagepub.com/doi/abs/10.1177/0272989X11435173 (Accessed: 15 March 2021).

Drost, E. A. (2011) 'Validity and Reliability in Social Science Research', *International Perspectives on Higher Education Research*, 38(1), pp. 105–124. Available at: http://dx.doi.org/ (Accessed: 12 March 2021).

Eastwood, J. R. *et al*. (2018) 'Increasing the accuracy and precision of relative telomere length estimates by RT qPCR', *Molecular ecology resources*, 18(1), pp. 68–78. doi: 10.1111/1755-0998.12711.

Eichorn, D. H. (1968) 'Variations in Growth Rate', *Childhood Education*, 44(5), pp. 286–291. doi: 10.1080/00094056.1968.10729296.

Enkel, T. *et al*. (2009) 'Ambiguous-Cue Interpretation is Biased Under Stress- and Depression-Like States in Rats', *Neuropsychopharmacology: official publication of the American College of Neuropsychopharmacology*, 35(4), pp. 1008–1015. doi: 10.1038/npp.2009.204.

'Environmental enrichment induces optimistic cognitive bias in rats' (2011) *Animal behaviour*, 81(1), pp. 169–175. doi: 10.1016/j.anbehav.2010.09.030.

European Comission (2015) *Development, integration and dissemination of animal-based welfare indicators, including pain, in commercially important husbandry species, with special emphasis on small ruminants, equidae & turkeys*. Available at: https://cordis.europa.eu/project/id/266213/reporting (Accessed: 15 March 2021).

'Evaluation of a novel feather scoring system for monitoring feather damaging behaviour in parrots' (2013) *The Veterinary Journal*, 196(2), pp. 247–252. doi: 10.1016/j.tvjl.2012.08.020.

Eysenck, M. W., Mogg, K., May, J., Richards, A., & Mathews, A. (1991) 'Bias in interpretation of ambiguous sentences related to threat in anxiety', *Journal of Abnormal Psychology*, 100(2), pp. 144–150. Available at: https://psycnet.apa.org/doiLanding?doi=10.1037/0021-843X.100.2.144 (Accessed: 15 March 2021).

Fraser, D. (1997) 'Preference and Motivation Testing', *n M.C. Appleby & B.O. Hughes (Eds.) Animal Welfare. New York: CAB Internationa*, pp. 159–173. Available at: https://www.wellbeingintlstudiesrepository.org/cgi/viewcontent.cgi?article=1001&context=valaexp (Accessed: 12 March 2021).

Fraser, D. (2008) 'Understanding animal welfare', *Acta veterinaria Scandinavica*, 50, pp. S1–12. doi: 10.1186/1751-0147-50-s1-s1.

Fureix, C. *et al*. (2012) 'Towards an Ethological Animal Model of Depression? A Study on Horses', *PloS one*, 7(6), p. e39280. doi: 10.1371/journal.pone.0039280.

Garner, J. P. (2005) ‘Stereotypies and Other Abnormal Repetitive Behaviors: Potential Impact on Validity, Reliability, and Replicability of Scientific Outcomes’, *ILAR journal / National Research Council, Institute of Laboratory Animal Resources*, 46(2), pp. 106–117. doi: 10.1093/ilar.46.2.106.

Gerhard Manteuffel, Birger Puppe, Peter C Schön (2004) ‘Vocalization of farm animals as a measure of welfare’, *Applied animal behaviour science*, 88(1-2), pp. 163–182. doi: 10.1016/j.applanim.2004.02.012.

Gewin, V. (2013) *Not all species deteriorate with age*. doi: 10.1038/nature.2013.14322.

Gillian V. Pepper, M. B. A. D. N. (2018) ‘Telomeres as integrative markers of exposure to stress and adversity: a systematic review and meta-analysis’, *Royal Society Open Science*, 5(8). doi: 10.1098/rsos.180744.

GJ Mason, N. R. L. (2004) ‘Can’t Stop won't stop: is stereotypy a reliable animal welfare indicator?’, *Animal Welfare*, 13, pp. 57–69. Available at: https://atrium.lib.uoguelph.ca/xmlui/bitstream/handle/10214/4716/Mason_%26_Latham_2004.pdf (Accessed: 15 March 2021).

Gotlib, I. H. and Krasnoperova, E. (1998) ‘Biased information processing as a vulnerability factor for depression’, *Behavior therapy*, 29(4), pp. 603–617. doi: 10.1016/S0005-7894(98)80020-8.

Gouveia, V. V. *et al*. (2008) ‘Life Satisfaction in Brazil: Testing the Psychometric Properties of the Satisfaction With Life Scale (SWLS) in Five Brazilian Samples’, *Social indicators research*, 90(2), pp. 267–277. doi: 10.1007/s11205-008-9257-0.

Grandin, T. (1998) ‘The feasibility of using vocalization scoring as an indicator of poor welfare during cattle slaughter’ (1998) Applied animal behaviour science, 56(2-4), pp. 121–128. doi: 10.1016/S0168-1591(97)00102-0.

Gregory A. Bryant, A. A. (2014) ‘The animal nature of spontaneous human laughter’, *Evolution and human behavior: official journal of the Human Behavior and Evolution Society*, 35(4), pp. 327–335. doi: 10.1016/j.evolhumbehav.2014.03.003.

Guesgen, M. J. *et al*. (2016) ‘Lambs show changes in ear posture when experiencing pain’. doi: 10.7120/09627286.25.2.171.

Guijt, A. M., Sluiter, J. K. and Frings-Dresen, M. H. W. (2007) ‘Test-retest reliability of heart rate variability and respiration rate at rest and during light physical activity in normal subjects’, *Archives of medical research*, 38(1), pp. 113–120. doi: 10.1016/j.arcmed.2006.07.009.

Hameed, R. H. (2018) 'Psychological and Biological Effects of Stress on Human Immune System', *Indian Journal of Public Health Research Development*, 9(8), pp. 184–189. doi: 10.5958/0976-5506.2018.00899.9 .

Hammond, P. J. (1991) 'Interpersonal comparisons of utility: Why and how they are and should be made', in Elster, J. and Roemer, J. E. (eds) *Interpersonal Comparisons of Well-Being*. Cambridge: Cambridge University Press, pp. 200–254. doi: 10.1017/cbo9781139172387.008.

Harrod, R. F. (1938) 'Scope and Method of Economics', *The Economic journal of Nepal*, 48(191), pp. 383–412. doi: 10.2307/2225434.

Helmut Segner, Henrik Sundh, Kurt Buchmann, Jessica Douxfils, Kristina Snuttan Sundell, Ce´dric Mathieu, Neil Ruane • Fredrik Jutfelt, Hilde Toften, Lloyd Vaughan (2012) 'Health of farmed fish: its relation to fish welfare and its utility as welfare indicator', *Fish Physiol Biochem*, 38, pp. 85–105. doi: 10.1007/s10695-011-9517-9.

Hemsworth, P. H. *et al*. (2011) 'Human–animal interactions at abattoirs: Relationships between handling and animal stress in sheep and cattle', *Applied animal behaviour science*, 135(1-2), pp. 24–33. doi: 10.1016/j.applanim.2011.09.007.

Hemsworth, P. H., Barnett, J. L. and Hansen, C. (1981) 'The influence of handling by humans on the behavior, growth, and corticosteroids in the juvenile female pig', *Hormones and behavior*, 15(4), pp. 396–403. doi: 10.1016/0018-506x(81)90004-0.

Hirschenhauser, K. *et al*. (2012) 'Excreted corticosterone metabolites differ between two galliform species, Japanese Quail and Chicken, between sexes and between urine and faecal parts of droppings', *Journal of Ornithology*, pp. 1179–1188. doi: 10.1007/s10336-012-0848-9.

'Inter-observer reliability testing of pig welfare outcome measures proposed for inclusion within farm assurance schemes' (2011) *The Veterinary Journal*, 190(2), pp. e100–e109. doi: 10.1016/j.tvjl.2011.01.012.

Is it better to be a wild rat or a factory farmed cow? A systematic method for comparing animal welfare (no date). Available at: http://www.charityentrepreneurship.com/1/post/2018/09/is-it-better-to-be-a-wild-rat-or-a-factory-farmed-cow-a-systematic-method-for-comparing-animal-welfare.html (Accessed: 15 March 2021).

Jafari, M. J. *et al*. (2019) 'The effects of combined exposure to noise and heat on human salivary cortisol and blood pressure', *International journal of occupational safety and ergonomics: JOSE*, pp. 1–9. doi: 10.1080/10803548.2019.1659578.

Järvi, T. (1989) ‘Synergistic effect on mortality in Atlantic salmon, Salmo salar , smolt caused by osmotic stress and presence of predators’, *Environmental biology of fishes*, 26(2), pp. 149–152. doi: 10.1007/BF00001031.

J. Beloor, H. K. Kang, Y. J. Kim, V. K. Subramani, I. S. Jang, S. H. Sohn and Y. S. Moon (2010) ‘The Effect of Stocking Density on Stress Related Genes and Telomeric Length in Broiler Chickens’, *Asian-Aust. J. Anim*, 23(4), pp. 437–443. Available at: https://www.ajas.info/upload/pdf/23-59.pdf (Accessed: 15 March 2021).

J. L. Barnett, P. H. H. (1990) ‘The validity of physiological and behavioural measures of animal welfare’, *Applied animal behaviour science*, 25(1-2), pp. 177–187. doi: 10.1016/0168-1591(90)90079-S.

Joel McGuire, A. M. B.-M. A. C. K. (2020) ‘Cash transfers’. Available at: https://www.happierlivesinstitute.org/cash-transfers.html (Accessed: 12 March 2021).

Jon M Watts, J. M. S. (1999) ‘Effects of restraint and branding on rates and acoustic parameters of vocalization in beef cattle’, *Applied animal behaviour science*, 62(2-3), pp. 125–135. doi: 10.1016/S0168-1591(98)00222-6.

Jovanović, V. (2016) ‘The validity of the Satisfaction with Life Scale in adolescents and a comparison with single-item life satisfaction measures: a preliminary study’, *Quality of life research: an international journal of quality of life aspects of treatment, care and rehabilitation*, 25(12), pp. 3173–3180. doi: 10.1007/s11136-016-1331-5.

Kaminitz, S. C. (2018) ‘Happiness Studies and the Problem of Interpersonal Comparisons of Satisfaction: Two Histories, Three Approaches’, *Journal of happiness studies*, 19(2), pp. 423–442. doi: 10.1007/s10902-016-9829-7.

Karasik, D. *et al*. (2004) ‘Genetic Contribution to Biological Aging: The Framingham Study’, *The journals of gerontology. Series A, Biological sciences and medical sciences*, 59(3), pp. B218–B226. doi: 10.1093/gerona/59.3.B218.

Keeling, L. (no date) ‘Definition of criteria for overall assessment of animal welfare’, *Animal welfare* . Available at: https://www.academia.edu/14868239/Definition_of_criteria_for_overall_assessment_of_animal_welfare (Accessed: 15 March 2021).

‘Kicking, rocking, and waving: Contextual analysis of rhythmical stereotypies in normal human infants’ (1981) *Animal behaviour*, 29(1), pp. 3–11. doi: 10.1016/S0003-3472(81)80146-7.

Kim, S. *et al*. (2011) 'Reliability and Short-Term Intra-Individual Variability of Telomere Length Measurement Using Monochrome Multiplexing Quantitative PCR', *PloS one*, 6(9), p. e25774. doi: 10.1371/journal.pone.0025774.

Kirkden, R. D. and Pajor, E. A. (2006) 'Using preference, motivation and aversion tests to ask scientific questions about animals' feelings', *Applied animal behaviour science*, 100(1), pp. 29–47. doi: 10.1016/j.applanim.2006.04.009.

Kotrschal, A., Ilmonen, P. and Penn, D. J. (2007) 'Stress impacts telomere dynamics', *Biology letters*, 3(2), p. 128. doi: 10.1098/rsbl.2006.0594.

Krantz, S., & Hammen, C. L (1979) 'Assessment of cognitive bias in depression', *Journal of Abnormal Psychology*, 88(6), pp. 611–619. Available at: https://psycnet.apa.org/record/1980-04473-001 (Accessed: 15 March 2021).

Kringelbach, M. L. (2010) 'The hedonic brain: A functional neuroanatomy of human pleasure', in Kringelbach, M. L. (ed.) *Pleasures of the brain , (pp*. New York, NY, US: Oxford University Press, viii, pp. 202–221. Available at: https://psycnet.apa.org/fulltext/2009-13385-012.pdf.

Kristensen T N, L. P. (2006) 'Physiological responses to heat stress and their potential use as indicators of reduced animal welfare in Jersey calves'. Available at: https://europepmc.org/article/cba/623602.

Kristiansen, T. S. *et al*. (2004) 'Swimming behaviour as an indicator of low growth rate and impaired welfare in Atlantic halibut (Hippoglossus hippoglossus L.) reared at three stocking densities'. Available at: https://pubag.nal.usda.gov/catalog/673553 (Accessed: 15 March 2021).

Krzysztof Wojtas, P. P. C. A. R. K. (2015) 'Cognitive bias test as a tool for accessing fish welfare', *Front. Mar. Sci. Conference Abstract: XV European Congress of Ichthyology*. Available at: https://www.frontiersin.org/10.3389/conf.FMARS.2015.03.00200/event_abstract (Accessed: 15 March 2021).

Lane, J. (2006) 'Can non-invasive glucocorticoid measures be used as reliable indicators of stress in animals?', *Animal welfare* , 15(4), pp. 331–342. Available at: https://www.ingentaconnect.com/content/ufaw/aw/2006/00000015/00000004/art00003.

Lee, H. M., Kim, K. S. and Lee, J. G. (2003) 'Investigation of Abnormal Eggs and Cortisol Stress Hormone in Laying Hens due to the Artificial Noise', *Journal of Korean Society of Environmental Engineers*, 25(7), pp. 13–860. Available at: https://www.koreascience.or.kr/article/JAKO200316240803462.page (Accessed: 12 March 2021).

Lewis, N. J. and Hurnik, J. F. (1990) 'Locomotion of Broiler Chickens in Floor Pens', *Poultry science*, 69(7), pp. 1087–1093. doi: 10.3382/ps.0691087.

Liu, Z., Torrey, S., Newberry, R. C., & Widowski, T (2020) 'Play behaviour reduced by environmental enrichment in fast-growing broiler chickens', *Applied animal behaviour science*, 232. doi: 10.1016/j.applanim.2020.105098.

Lucy Asher, Mary Friel, Kym Griffin and Lisa M. Collins (2016) 'Mood and personality interact to determine cognitive biases in pigs', *Biology Letters*, 12(11). doi: 10.1098/rsbl.2016.0402.

Magnus, K., Diener, E., Fujita, F., & Pavot, W (1993) 'Extraversion and neuroticism as predictors of objective life events: A longitudinal analysis', *Journal of Personality and Social Psychology*, 65(5), pp. 1046–1053. doi: 10.1037/0022-3514.65.5.1046.

Main, D. (2018) 'Why Koko the Gorilla Mattered', *National Geographic*, 21 June. Available at: https://www.nationalgeographic.com/animals/article/gorillas-koko-sign-language-culture-animals (Accessed: 12 March 2021).

Manal A. Fouad, A. H. A. R. A. E. S. M. B. (2008) 'Broilers Welfare and Economics under Two Management Alternatives on Commercial Scale', *International Journal of Poultry Science*, 7(12), pp. 1167–1173. doi: 10.3923/ijps.2008.1167.1173.

Marek Navrátil, C. A. L. (2006) 'Temporal Stability of the Czech Translation of the Satisfaction with Life Scale: Test-Retest Data over One Week', *NAVRÁTIL, MAREK; LEWIS, CHRISTOPHER ALAN (2006). TEMPORAL STABILITY OF THE CZECH TRANSLATION OF THE SATISFACTION WITH LIFE SCALE: TEST-RETEST DATA OVER ONE WEEK 1. Psychological Reports, 98(3), 918–920. doi:10.2466/pr0.98.3.918-920*, 98(3), pp. 918–920. Available at: https://journals.sagepub.com/doi/abs/10.2466/pr0.98.3.918-920 (Accessed: 15 March 2021).

Marshall, G. N., Wortman, C. B., Kusulas, J. W., Hervig, L. K., & Vickers, R. R., Jr. (1992) 'Distinguishing optimism from pessimism: Relations to fundamental dimensions of mood and personality', *Journal of Personality and Social Psychology*, 62(6), pp. 1067–1074. doi: 10.1037/0022-3514.62.6.1067.

Martín-María, N. M. S. *et al*. (2017) 'The Impact of Subjective Well-being on Mortality: A Meta-Analysis of Longitudinal Studies in the General Populatio', *Psychosomatic Medicine*, 79(5), pp. 565–575. doi: 10.1097/PSY.0000000000000444.

Martins, C. I. M. *et al*. (2011) 'Behavioural indicators of welfare in farmed fish', *Fish physiology and biochemistry*, 38(1), pp. 17–41. doi: 10.1007/s10695-011-9518-8.

Mason, G. J. (1993) ‘Age and Context Affect the Stereotypies of Caged Mink’, *Behaviour*, 127(3-4), pp. 191–229. doi: 10.1163/156853993X00029.

McCulloch, S. P. (2012) ‘A Critique of FAWC’s Five Freedoms as a Framework for the Analysis of Animal Welfare’, *Journal of agricultural & environmental ethics*, 26(5), pp. 959–975. doi: 10.1007/s10806-012-9434-7.

McKenzie, D. J. *et al*. (2012) ‘Effects of stocking density and sustained aerobic exercise on growth, energetics and welfare of rainbow trout’, *Aquaculture* , 338-341, pp. 216–222. doi: 10.1016/j.aquaculture.2012.01.020.

McMillan, F. D. (2003) ‘Maximizing Quality of Life in III Animals’, *Journal of the American Animal Hospital Association*, 39(3), pp. 227–235. doi: 10.5326/0390227.

Md. Mahmudul Hasan Sagar and A. K. M. Rezaul Karim (2014) ‘The psychometric properties of Satisfaction With Life Scale for police population in Bangladeshi Culture’, *The International Journal of Social Sciences*, 28(1). Available at: https://www.researchgate.net/publication/269036647_The_psychometric_properties_of_Satisfaction_With_Life_Scale_for_police_population_in_Bangladeshi_Culture.

Mellor, D. J. (2016) ‘Moving beyond the “Five Freedoms” by Updating the “Five Provisions” and Introducing Aligned “Animal Welfare Aims”’, *Animals : an Open Access Journal from MDPI*, 6(10). doi: 10.3390/ani6100059.

Mellor, D. J. (2017) ‘Operational Details of the Five Domains Model and Its Key Applications to the Assessment and Management of Animal Welfare’, *Animals : an Open Access Journal from MDPI*, 7(8). doi: 10.3390/ani7080060.

Michael Mendl, Oliver H.P. Burman, Richard M.A. Parker, Elizabeth S. Paul (2009) ‘Cognitive bias as an indicator of animal emotion and welfare: Emerging evidence and underlying mechanisms’, *Applied Animal Behaviour Science*, 118, pp. 161–181. Available at: https://www.researchgate.net/profile/Oliver-Burman/publication/240398413_Cognitive_bias_as_an_indicator_of_animal_emotion_and_welfare_Emerging_evidence_and_underlying_mechanisms/links/59e47b31458515393d60e8c4/Cognitive-bias-as-an-indicator-of-animal-emotion-and-welfare-Emerging-evidence-and-underlying-mechanisms.pdf.

Michael Mendl , Oliver H.P. Burman, Richard M.A. Parker, Elizabeth S. Paul (2009) ‘Cognitive bias as an indicator of animal emotion and welfare: Emerging evidence and underlying mechanisms’, *Applied Animal Behaviour Science*, 118, pp. 161–181. Available at: https://www.researchgate.net/profile/Oliver_Burman/publication/240398413_Cognitive_bias_as_an_indicator_of_animal_emotion_and_welfare_Emerging_evidence_and_underlying_mechanisms/links/59e47b31458515393d60e8c4/Cognitive-bias-as-

an-indicator-of-animal-emotion-and-welfare-Emerging-evidence-and-underlying-mechanisms.pdf (Accessed: 15 March 2021).

Michael P. Mcloughlin, R. S. A. A. G. M. (2019) 'Automated bioacoustics: methods in ecology and conservation and their potential for animal welfare monitoring', *Journal of The Royal Society Interface*, 16(155). doi: 10.1098/rsif.2019.0225.

MichaelStJules's Shortform (no date). Available at: https://forum.effectivealtruism.org/posts/GK7Qq4kww5D8ndckR/michaelstjules-s-shortform?commentId=LZNATg5BoBT3w5AYz (Accessed: 12 March 2021).

Miller, L. J. *et al*. (2020) 'Behavioral Diversity as a Potential Indicator of Positive Animal Welfare', *Animals*, 10(7), p. 1211. doi: 10.3390/ani10071211.

Mintline, E. M. *et al*. (2013) 'Play behavior as an indicator of animal welfare: Disbudding in dairy calves', *Applied animal behaviour science*, 144(1-2), pp. 22–30. doi: 10.1016/j.applanim.2012.12.008.

Miranda-de la Lama, G. C. *et al*. (2010) 'Effect of the pre-slaughter logistic chain on some indicators of welfare in lambs', *Livestock science*, 128(1), pp. 52–59. doi: 10.1016/j.livsci.2009.10.013.

Moberg, G. P. and Mench, J. A. (2000) *The Biology of Animal Stress: Basic Principles and Implications for Animal Welfare*. CABI. Available at: https://play.google.com/store/books/details?id=LmKCN-7kluYC.

'Mood and the speed of decisions about anticipated resources and hazards' (2011) *Evolution and human behavior: official journal of the Human Behavior and Evolution Society*, 32(1), pp. 21–28. doi: 10.1016/j.evolhumbehav.2010.07.005.

Morrison, S. R. (1983) 'Ruminant heat stress: effect on production and means of alleviation', *Journal of animal science*, 57(6), pp. 1594–1600. doi: 10.2527/jas1983.5761594x.

Muneer, M. A. *et al*. (1988) 'Immunosuppression in animals', *The British veterinary journal*, 144(3), pp. 288–301. doi: 10.1016/0007-1935(88)90116-9.

Nephew, B. C., Kahn, S. A. and Romero, L. M. (2003) 'Heart rate and behavior are regulated independently of corticosterone following diverse acute stressors', *General and comparative endocrinology*, 133(2), pp. 173–180. doi: 10.1016/s0016-6480(03)00165-5.

Neto, F. (1993) 'The satisfaction with life scale: Psychometrics properties in an adolescent sample', *Journal of youth and adolescence*, 22(2), pp. 125–134. doi: 10.1007/BF01536648.

‘Newborn and 5-week-old calves vocalize in response to milk deprivation’ (2001) *Applied animal behaviour science*, 74(3), pp. 165–173. doi: 10.1016/S0168-1591(01)00164-2.

Nicol, C. J. *et al*. (2009) ‘Associations between welfare indicators and environmental choice in laying hens’, *Animal behaviour*, 78(2), pp. 413–424. doi: 10.1016/j.anbehav.2009.05.016.

Niemiec, C. P. (2014) ‘Eudaimonic Well-Being’, in Michalos, A. C. (ed.) *Encyclopedia of Quality of Life and Well-Being Research*. Dordrecht: Springer Netherlands, pp. 2004–2005. doi: 10.1007/978-94-007-0753-5_929.

Norberg, M. M. *et al*. (2008) ‘Quality of life in obsessive-compulsive disorder: an evaluation of impairment and a preliminary analysis of the ameliorating effects of treatment’, *Depression and Anxiety*, pp. 248–259. doi: 10.1002/da.20298.

Palme, R. (2012) ‘Monitoring stress hormone metabolites as a useful, non-invasive tool for welfare assessment in farm animals’, *Animal welfare* , 21(3), pp. 331–337. doi: 10.7120/09627286.21.3.331.

P.-A. Morin, Y. Chorfi, J. Dubuc,J.-P. Roy,D. Santschi, S. Dufour (2017) ‘Short communication: An observational study investigating inter-observer agreement for variation over time of body condition score in dairy cows’, *Journal of dairy science*, 100(4), pp. 3086–3090. doi: 10.3168/jds.2016-11872.

Panksepp, J. and Burgdorf, J. (2003) ‘“Laughing” rats and the evolutionary antecedents of human joy?’, *Physiology & behavior*, 79(3). doi: 10.1016/s0031-9384(03)00159-8.

Part, C. E. *et al*. (2016) ‘Prevalence rates of health and welfare conditions in broiler chickens change with weather in a temperate climate’, *Royal Society Open Science*, 3(9). doi: 10.1098/rsos.160197.

Pauliny, A. *et al*. (2015) ‘Rapid growth accelerates telomere attrition in a transgenic fish’, *BMC evolutionary biology*, 15(1), pp. 1–10. doi: 10.1186/s12862-015-0436-8.

Pavot, W. *et al*. (1991) ‘Further validation of the Satisfaction with Life Scale: evidence for the cross-method convergence of well-being measures’, *Journal of personality assessment*, 57(1), pp. 149–161. doi: 10.1207/s15327752jpa5701_17.

Paweł A. Atroszko – Artur Sawicki – Aleksandra Mąkinia – Bartosz Atroszko (2017) ‘FURTHER VALIDATION OF SINGLE-ITEM SELF-REPORT MEASURE OF SATISFACTION WITH LIFE’, *7th Comparative European Research*, pp. 107–110. Available at: https://is.muni.cz/repo/1391745/Svarcova_sbornik.pdf#page=107.

Pereira, D. F. *et al*. (2013) ‘Machine vision to identify broiler breeder behavior’, *Computers and Electronics in Agriculture*, 99, pp. 194–199. doi: 10.1016/j.compag.2013.09.012.

Petrik, M. T., Guerin, M. T. and Widowski, T. M. (2013) 'Keel fracture assessment of laying hens by palpation: inter-observer reliability and accuracy', *The Veterinary record*, 173(20), p. 500. doi: 10.1136/vr.101934.

Phythian, C. *et al*. (2013) 'Inter-observer reliability of Qualitative Behavioural Assessments of sheep', *Applied animal behaviour science*, 144(1-2), pp. 73–79. doi: 10.1016/j.applanim.2012.11.011.

Phythian, C. J. *et al*. (2012) 'Reliability of indicators of sheep welfare assessed by a group observation method', *The Veterinary Journal*, 193(1), pp. 257–263. doi: 10.1016/j.tvjl.2011.12.006.

Pohle, K. and Cheng, H.-W. (2009) 'Furnished cage system and hen well-being: Comparative effects of furnished cages and battery cages on behavioral exhibitions in White Leghorn chickens', *Poultry science*, 88(8), pp. 1559–1564. doi: 10.3382/ps.2009-00045.

Puterman, E. *et al*. (2015) 'Determinants of telomere attrition over one year in healthy older women: Stress and health behaviors matter', *Molecular psychiatry*, 20(4), p. 529. doi: 10.1038/mp.2014.70.

Qualitative Behaviour Assessment (no date). Available at: https://warwick.ac.uk/fac/soc/sociology/research/currentresearch/interspeciesconnectedness/summary/qba/ (Accessed: 15 March 2021).

Quality®, W. (no date) *Welfare Quality® assessment protocol for cattle*. *Welfare Quality® Consortium, Lelystad, Netherlands*. Available at: http://www.welfarequalitynetwork.net/media/1088/cattle_protocol_without_veal_calves.pdf (Accessed: 15 March 2021).

Ramsay, J. M. *et al*. (2009) 'Whole-body cortisol response of zebrafish to acute net handling stress', *Aquaculture* , 297(1-4), pp. 157–162. doi: 10.1016/j.aquaculture.2009.08.035.

Reefmann, N. *et al*. (2009) 'Ear and tail postures as indicators of emotional valence in sheep', *Applied animal behaviour science*, 118(s 3–4), pp. 199–207. doi: 10.1016/j.applanim.2009.02.013.

'Release from restraint generates a positive judgement bias in sheep' (2010) *Applied animal behaviour science*, 122(1), pp. 28–34. doi: 10.1016/j.applanim.2009.11.003.

Rice, C. M. (2013) 'Defending the objective list theory of well-being: Defending the objective list theory of well-being', *Ratio*, 26(2), pp. 196–211. doi: 10.1111/rati.12007.

RICHARD J. DAVIDSON AND BRIANNA S. SCHUYLER (no date) 'NEUROSCIENCE OF HAPPINESS'. Available at: https://happyecho.com/wp-content/uploads/2015/05/WHR15.pdf#page=90.

Robin Technologies, Inc. http://www. robintek.com (no date) *Videomex-ONE Video Tracking System*. Available at: http://www.colinst.com/products/animal-activity-meter-videomex-one (Accessed: 15 March 2021).

Rosengren, L. *et al*. (2015) 'Psychometric properties of the Satisfaction With Life Scale in Parkinson's disease', *Acta neurologica Scandinavica*, 132(3), pp. 164–170. doi: 10.1111/ane.12380.

Rushen, J. (1986) 'Aversion of sheep for handling treatments: Paired-choice studies', *Applied animal behaviour science*, 16(4), pp. 363–370. doi: 10.1016/0168-1591(86)90008-0.

Rutherford, K. M. *et al*. (2012) 'Qualitative Behavioural Assessment of emotionality in pigs', *Applied animal behaviour science*, 139(3-4). doi: 10.1016/j.applanim.2012.04.004.

Sartory, G., Rachman, S. and Grey, S. (1977) 'An investigation of the relation between reported fear and heart rate', *Behaviour research and therapy*, 15(5), pp. 435–438. doi: 10.1016/0005-7967(77)90048-1.

Schukraft, J. (2020) *Comparisons of Capacity for Welfare and Moral Status Across Species — Rethink Priorities*. Rethink Priorities. Available at: https://www.rethinkpriorities.org/blog/2020/5/16/comparisons-of-capacity-for-welfare-and-moral-status-across-species (Accessed: 15 March 2021).

Schutte, N. S. and Malouff, J. M. (2015) 'The association between depression and leukocyte telomere length: a meta-analysis', *Depression and anxiety*, 32(4), pp. 229–238. doi: 10.1002/da.22351.

Segerstrom, S. C. and Miller, G. E. (2004) 'Psychological stress and the human immune system: a meta-analytic study of 30 years of inquiry', *Psychological bulletin*, 130(4), pp. 601–630. doi: 10.1037/0033-2909.130.4.601.

Sénèque, E. *et al*. (2019) 'Could posture reflect welfare state? A study using geometric morphometrics in riding school horses', *PloS one*, 14(2), p. e0211852. doi: 10.1371/journal.pone.0211852.

Shields, S. and Greger, M. (2013) 'Animal Welfare and Food Safety Aspects of Confining Broiler Chickens to Cages', *Animals*, 3(2), pp. 386–400. doi: 10.3390/ani3020386.

Short, S. J. *et al*. (2016) 'Correspondence between hair cortisol concentrations and 30-day integrated daily salivary and weekly urinary cortisol measures', *Psychoneuroendocrinology*, 71, pp. 12–18. doi: 10.1016/j.psyneuen.2016.05.007.

Smyth, J. M. *et al*. (2017) 'Global life satisfaction predicts ambulatory affect, stress, and cortisol in daily life in working adults', *Journal of behavioral medicine*, 40(2), pp. 320–331. doi: 10.1007/s10865-016-9790-2.

Social Behavior in Farm Animals (no date). Available at: https://books.google.com/books/about/Social_Behavior_in_Farm_Animals.html?id=DZSXTHrurzcC (Accessed: 15 March 2021).

Spruijt, B. M., van Hooff, J. A. and Gispen, W. H. (1992) 'Ethology and neurobiology of grooming behavior', *Physiological reviews*, 72(3). doi: 10.1152/physrev.1992.72.3.825.

Steffens, A. B. (1969) 'Rapid absorption of glucose in the intestinal tract of the rat after ingestion of a meal', *Physiology & behavior*, 4(5), pp. 829–832. doi: 10.1016/0031-9384(69)90125-5.

Stephens, R., Atkins, J. and Kingston, A. (2009) 'Swearing as a response to pain', *Swearing as a response to pain. NeuroReport*, 20(12), pp. 1056–1060. Available at: https://journals.lww.com/neuroreport/Abstract/2009/08050/Swearing_as_a_response_to_pain.4.aspx (Accessed: 15 March 2021).

Stockman, C. A., Collins, T., Barnes, A. L., Miller, D., Wickham, S. L., Beatty, D. T., ... & Fleming, P. A (2011) 'Qualitative Behavioural Assessment and Quantitative Physiological Measurement of Cattle Naïve and Habituated to Road Transport', *Animal Production Science*, 51(3), pp. 240–249. Available at: https://www.wellbeingintlstudiesrepository.org/cgi/viewcontent.cgi?article=1052&context=acwp_vsm (Accessed: 15 March 2021).

Tang, M. and Boisclair, D. (2011) 'Relationship between respiration rate of juvenile brook trout (Salvelinus fontinalis), water temperature, and swimming characteristics', *Canadian journal of fisheries and aquatic sciences. Journal canadien des sciences halieutiques et aquatiques*, 52(10), pp. 2138–2145. doi: 10.1139/f95-806.

Taylor, K. D. and Mills, D. S. (2007) 'Is quality of life a useful concept for companion animals?' Available at: https://www.ingentaconnect.com/content/ufaw/aw/2007/00000016/a00102s1/art00009 (Accessed: 15 March 2021).

Teng, K. T.-Y. *et al*. (2018) 'Welfare-Adjusted Life Years (WALY): A novel metric of animal welfare that combines the impacts of impaired welfare and abbreviated lifespan', *PloS one*, 13(9), p. e0202580. doi: 10.1371/journal.pone.0202580.

The Five Freedoms for animals (no date). Available at: https://www.animalhumanesociety.org/health/five-freedoms-animals (Accessed: 15 March 2021).

Thelen, E. (1979) 'Rhythmical stereotypies in normal human infants', *Animal behaviour*, 27(Pt 3), pp. 699–715. doi: 10.1016/0003-3472(79)90006-x.

Thompson, T. M. *et al*. (2012) 'Comparison of Whole-Genome DNA Methylation Patterns in Whole Blood, Saliva, and Lymphoblastoid Cell Lines', *Behavior genetics*, 43(2), pp. 168–176. doi: 10.1007/s10519-012-9579-1.

‘Threat is in the eye of the beholder: Social anxiety and the interpretation of ambiguous facial expressions’ (2007) *Behaviour research and therapy*, 45(4), pp. 839–847. doi: 10.1016/j.brat.2006.05.004.

Trimmer, P. C. *et al*. (2013) ‘On the evolution and optimality of mood States’, *Behavioral sciences*, 3(3), pp. 501–521. doi: 10.3390/bs3030501.

Trude Arnesen, M. T. (2004) ‘Roughly right or precisely wrong? Systematic review of quality-of-life weights elicited with the time trade-off method’, *Journal of Health Services Research & Policy*, 9(1). Available at: https://journals.sagepub.com/doi/abs/10.1258/135581904322716111 (Accessed: 15 March 2021).

T Seo, K Date, T Daigo, F Kashiwamura†and S Sato (2007) ‘Welfare assessment of Japanese dairy farms using the Animal Need Index’, *Animal Welfare*, 16, pp. 211–223. Available at: https://www.researchgate.net/publication/233577925_Welfare_assessment_of_Japanese_dairy_farms_using_the_Animal_Need_Index (Accessed: 15 March 2021).

Vassar, M. (2007) ‘A note on the score reliability for the Satisfaction With Life Scale: an RG study’, *Social indicators research*, 86(1), pp. 47–57. doi: 10.1007/s11205-007-9113-7.

Verkerk, G. A. *et al*. (1998) ‘Characterization of Milk Cortisol Concentrations as a Measure of Short-Term Stress Responses in Lactating Dairy Cows’, *Animal welfare* , 7(1), pp. 77–86. Available at: https://www.ingentaconnect.com/content/ufaw/aw/1998/00000007/00000001/art00007.

Vestergaard, K. S. *et al*. (1999) ‘Regulation of dustbathing in feathered and featherless domestic chicks: the Lorenzian model revisited’, *Animal behaviour*, 58(5). doi: 10.1006/anbe.1999.1233.

Weeks, C. A., Lambton, S. L. and Williams, A. G. (2016) ‘Implications for Welfare, Productivity and Sustainability of the Variation in Reported Levels of Mortality for Laying Hen Flocks Kept in Different Housing Systems: A Meta-Analysis of Ten Studies’, *PloS one*, 11(1), p. e0146394. doi: 10.1371/journal.pone.0146394.

Welfare Quality Network (no date). Available at: http://www.welfarequalitynetwork.net/en-us/reports/assessment-protocols/ (Accessed: 15 March 2021).

Wemelsfelder, F. *et al*. (2012) ‘Assessing pig body language: Agreement and consistency between pig farmers, veterinarians, and animal activists1’, *Journal of animal science*, 90(10), pp. 3652–3665. doi: 10.2527/jas.2011-4691.

Wemelsfelder, F., Hunter, T. E., Mendl, M. T., & Lawrence, A. B. ((2001) ‘Assessing the “Whole Animal”: A Free Choice Profiling Approach’, *Animal Behaviour*, 62 (2), pp. 209–

220. Available at: https://www.wellbeingintlstudiesrepository.org/cgi/viewcontent.cgi?article=1096&context=acwp_asie (Accessed: 15 March 2021).

West, J. *et al*. (2004) 'Effects of Hatha yoga and African dance on perceived stress, affect, and salivary cortisol', *Annals of behavioral medicine: a publication of the Society of Behavioral Medicine*, 28(2), pp. 114–118. doi: 10.1207/s15324796abm2802_6.

Wickham, S. L. *et al*. (2012) 'Qualitative behavioral assessment of transport-naive and transport-habituated sheep', *Journal of animal science*, 90(12). doi: 10.2527/jas.2010-3451.

Willen, C. W. &stefanie (2010) 'The Reliability and Repeatability of a Lameness Scoring System for Use as an Indicator of Welfare in Dairy Cattle', *Acta Agriculturae Scandinavica*, 51, pp. 103–107. Available at: https://www.tandfonline.com/doi/abs/10.1080/090647001316923162 (Accessed: 15 March 2021).

Willett, J. B. (1989) 'Some Results on Reliability for the Longitudinal Measurement of Change: Implications for the Design of Studies of Individual Growth', *Educational and psychological measurement*, 49(3), pp. 587–602. doi: 10.1177/001316448904900309.

Williams, E. *et al*. (2006) 'Associations between whole-blood serotonin and subjective mood in healthy male volunteers', *Biological psychology*, 71(2), pp. 171–174. doi: 10.1016/j.biopsycho.2005.03.002.

Wulfert, E. *et al*. (2005) 'Heart rate arousal and excitement in gambling: winners versus losers', *Psychology of addictive behaviors: journal of the Society of Psychologists in Addictive Behaviors*, 19(3), pp. 311–316. doi: 10.1037/0893-164X.19.3.311.

Yardley, J. K. and Rice, R. W. (1991) 'The relationship between mood and subjective well-being', *Social indicators research*, 24(1), pp. 101–111. doi: 10.1007/BF00292653.

Yon, L. *et al*. (2019) 'Development of a behavioural welfare assessment tool for routine use with captive elephants', *PloS one*, 14(2), p. e0210783. doi: 10.1371/journal.pone.0210783.

Yue, S., Moccia, R. D. and Duncan, I. J. H. (2004) 'Investigating fear in domestic rainbow trout, Oncorhynchus mykiss, using an avoidance learning task', *Applied animal behaviour science*, 87(3), pp. 343–354. doi: 10.1016/j.applanim.2004.01.004.

Image Credit: Jo-Anne McArthur