Welfare comparisons within and across species

Heather Browning[1,2] ⬤

Abstract
One of the biggest problems in applications of animal welfare science is our ability to make comparisons between different individuals, both within and across species. Although welfare science provides methods for measuring the welfare of individual animals, there's no established method for comparing measures between individuals. In this paper I diagnose this problem as one of underdetermination—there are multiple conclusions given the data, arising from two sources of variation that we cannot distinguish: variation in the underlying target variable (welfare experience) and in the relationship of measured indicators to the target. I then describe some of the possible methods of making comparisons, based on the use of similarity assumptions that will have greater or lesser justification in different circumstances, and the alternative methods we may use when direct comparisons are not possible. In the end, all our available options for making welfare comparisons are imperfect, and we need to make explicit context-specific decisions about which will be best for the task at hand while acknowledging their potential limitations. Future developments in our understanding of the biology of sentience will help strengthen our methods of making comparisons.

1 Introduction

Imagine you're a zoo manager, looking at proposals for spending this year's renovation budget. You could add new logs to a tank housing ten lungfish, which they'll enjoy as extra hiding and shelter places, or a heated rock to the lion exhibit, which the lion can use for warming up on cold days. When considering which option is

---

✉ Heather Browning
drheatherbrowning@gmail.com

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

2 Department of Philosophy, University of Southampton, Southampton, UK

🖄 Springer

best, your primary concern is the welfare of the animals in your charge; you want to spend the money in the way that provides the largest welfare increase. This decision requires a comparison of the benefits the different animals may experience as a result of the change: comparing the welfare gain to the lion to that of the ten lungfish. This is one example of an intersubjective welfare comparison; there are many other examples of such comparisons, made for a variety of reasons.

Perhaps the most common application is utilitarian decision-making over options including the utility of multiple species. Any construction of a social welfare function that includes multiple individuals, of the same or of different species, will require comparisons of relative welfare levels (Budolfson & Spears, 2020, Browning, 2022a). This includes decisions about how to prioritise charitable animal welfare interventions between those benefitting different species, as well as comparisons between human and nonhuman animal welfare, such as when performing cost–benefit analyses of practices that harm animals to benefit humans (e.g. medical experimentation) or contain potential harms or benefits to both (e.g. climate change policy). Similarly, we have decisions regarding resource allocation, like the example described above. Institutions (such as zoos) that hold multiple species must constantly make decisions about the distribution of resources between animals to achieve the best overall outcomes. Limitations to resources, such as money and husbandry time, create decisions about distribution under scarcity that will require making trade-offs between provision of benefits to different individuals or groups of animals.

Although interspecies comparisons may be the most common, or the most high-profile, we also need to make other types of comparisons within species. When making management decisions for the life of an animal—weighing the trade-offs required for making their lives go well overall—we make intrasubjective comparisons between past and future versions of the same individual. For example, whether we should put a young animal through a painful medical procedure to prevent health problems in later life, or cause frustration through denial of a favourite food type that could cause future obesity. These comparisons require comparative information about the degree of harm and benefit of these different actions for a single individual over time. Where we think the values or experiences of an individual will vary over time, such intrasubjective comparisons can be treated as a form of intersubjective comparison (Pettigrew, 2019).

Another type of comparison that is rarely discussed is performed within the practice of animal welfare science. Studying the welfare of animals under different conditions requires taking groups of animals and placing them under conditions such as different feeding regimes, environmental parameters, or social groupings. Measurement of behavioural and physiological indicators is then used to draw conclusions about the effects of these conditions on the welfare of the animals. Importantly, the tests are performed on small groups of animals, with results that are assumed to be relevant to other members of the species. Often, the different experimental conditions will be performed on different groups, and the results from each group compared. Here we have two ways in which intersubjective comparisons are necessary—in making comparisons between experimental groups and in extrapolating results to other members of the species; both of which will typically occur within species.

Throughout most of this paper I'll refer to interspecies comparisons as they are the most extreme case, but the problems described will also apply to intraspecies and even intraindividual comparisons and for simplicity, the examples I use will often be intraspecies.

As these examples show, we're constantly required to make welfare comparisons and they are relevant to many groups—consumers, policy-makers, activists, and ethicists. However, although we frequently make such comparisons, they are problematic, especially when performed unreflectively. In Section Two I'll describe the empirical problem of welfare comparisons and how it arises from an underdetermination of the conclusions from the available data. In Section Three I'll discuss the possible methods for making welfare comparisons, starting with a brief discussion of some proposed solutions to the comparisons problem in the human case and why they fail, before moving on to describe two methods of making welfare comparisons using background similarity assumptions, and when use of these is justified. In Section Four I will address the alternative methods that may be used when welfare comparisons are not justified, before moving on in Section Five to present my conclusions and recommend future research directions.

2 The problem of welfare comparisons

There are two problems that arise regarding welfare comparisons: an empirical problem and a moral problem. The empirical problem is the question of how we compare the measures of welfare between different species. That is, our ability to make empirical judgements about how much welfare a given animal is experiencing and to rank this against the experiences of others. The moral problem regards how we assign different moral weight to different species or individuals within our ethical decision-making. In some cases, particularly utilitarian frameworks, this depends almost entirely on the welfare capacity of the individual. However, we'll also have other reasons for setting moral weights independently of the level of welfare experience, such as our relationships with the individuals, or regarding the different features or capacities of individuals aside from that for welfare. I will not be addressing the moral problem in this paper, rather focussing on the empirical aspect of the comparison problem—looking at the comparative measurement of welfare rather than its application in ethics. I take it that the answers I provide here will be useful in informing ethical deliberation and in the final section, when looking at how we might move forwards in cases where the other methods are inapplicable, I'll briefly engage a little of the ethical theory.

In this paper, I will be assuming a hedonic, or subjective account of animal welfare, in which welfare consists in the integrated set of consciously experienced valenced mental states (sometimes called feelings, or affects). Though there is not space here to justify in depth the use of this concept (see Browning, 2020a for a full defence), there are a couple of reasons to prefer it. First, this is arguably the most commonly used concept within animal welfare science (e.g. Duncan, 2002; Mellor, 2016), and thus forms the basis of the most commonly encountered version of the problem of interspecies welfare comparisons. It is also one of the primary views

within animal ethics (e.g. Singer, 2016); the concept is typically adopted because of its ethical weight—the feelings of animals grounding our ethical concern for them. It is far less clear why we should care about the physical functioning or natural flourishing of an animal; in particular why and how this is differentiated from similar states in other organisms, such as plants.

Even for those who may not accept this view of welfare, this discussion is still relevant—the problems arise for any view that accepts that subjective experience comprises at least part of welfare and wants to have the means to compare strength of feeling between different animals. Other concepts of animal welfare (such as preferences, biological functioning, or natural living) are most often combined in a multi-component account alongside subjective experiencing (Fraser et al., 1997), and so their use could not entirely escape the fundamental problem of comparing strength of animal feelings. However, in Section Five I will also return to this issue and look briefly at how these alternative accounts of animal welfare might deal with the problem of welfare comparisons.

We are certainly able to measure the welfare of individual animals. As mentioned, in animal welfare science, scientists use different indicators, such as changes in behaviour or physiological variables, to measure the changes in welfare under different conditions. When measuring the welfare of any individual animal, we're quantifying its welfare—in some sense defining the 'units' of welfare for that individual.[1] We can establish both the full scale of welfare experience for an individual—its maximum and minimum levels and the associated indicator values—as well as where the individual sits along this scale at any particular time or under given conditions.

The question for welfare comparisons is, do different individuals have different scales? And if they do, how can we convert the 'units' of welfare between the scales? Having different scales of measurement is not in itself a problem—think of scales like measurement of length (centimetres and inches) and temperature (Celsius and Fahrenheit). We're able to convert measurements between the scales because we have the appropriate formulas for doing so. In welfare comparisons what we lack is the appropriate conversion formulas. We may have the measurements of welfare of one animal and those of another animal—both of which quantify welfare in relation to the scale for that animal—but we don't know how to convert units between the scales of each individual and compare them on a common scale, to be able to say something about how many measured units of welfare for one animal are equivalent to a unit of welfare for another.

It's entirely plausible that different individuals could experience vastly different levels of welfare, and that they do not reflect these differences in measurable indicators. We see versions of this in real-world situations: e.g. people can vary quite a lot with respect to pain thresholds and the degree to which they express pain reactions, making it very difficult to compare pain experience between individuals (Nielsen et al., 2005). It may be the case that some animals have reduced affect, where the

---

[1] Note that in this paper I'm assuming that subjective welfare is measurable on a ratio scale—see Browning (2022b) for a defence of this claim.

intensity of all their experiences may be small—their highs are not particularly high nor their lows particularly low. Others, by contrast, might be capable of reaching far higher heights and far deeper lows—their intensity is just greater overall. If such individuals exist without showing different indicator responses, then (as the underlying subjective states are private and inaccessible) we might never know whether or when they occur, and this undermines our ability to trust such comparisons. This goes beyond simply the problem of other minds. While the problem of other minds looks at our justification for inferring the presence of mental states in other individuals (human or animal); the problem of welfare comparisons goes beyond this to look specifically at the relative intensity of feeling within different minds. We are not concerned with similarities in the qualitative character of mental states, but rather direct comparisons of their intensity; which while similar in some regards, is difficult in a different way.

The reason it's difficult to make interspecies welfare comparisons is because we have a problem of underdetermination. That is, there are multiple possible conclusions that are compatible with our data. This occurs because there are two potential sources of variation that can explain our observed data. The first is variation in the values of the underlying target state (i.e. welfare experience). The second is variation in the relationship between the measured indicator and the target state. Some animals may be highly reactive, showing large changes in their measured indicators to only small increases or decreases in their subjective experience. Others may be more circumspect, showing only small external responses to large subjective changes. We have no way of testing for this possibility, and no a priori reason to rule it out. This is not just hypothetical scepticism—even within species, differences in individual behavioural and physiological responses to positive and negative stimuli are common (e.g. Boccia et al., 1995; Izzo et al., 2011; Manteca & Deag, 1994), and it's difficult in these cases to determine whether or not results imply a welfare difference.

As we don't know in any situation which of these types of variation are responsible for the results we observe, we are unable to draw justified conclusions regarding the comparative welfare of different individuals using this data alone. Under an observed difference in overall response, we don't know which of these factors—difference in level of welfare experience, or in indicator response—is responsible for this, or indeed if both are varying simultaneously. We can't simply infer intensity of feeling from observed behaviour, without some further justification for doing so (I will discuss some such potential justifications in Sect. 3.4). And without such information, we cannot make comparisons.

We can see how this works, using a brief example. Here, though I use an example of an *intra*species comparison, this is intended primarily for the sake of simplicity and also to illustrate how difficult the problem is even in its simplest form. As I will highlight throughout, the problem is even greater for the interspecies case in which we are primarily interested.

I used to work with two otters—Sneezy and Paddy. Imagine that each are given some yabbies,[2] and their behavioural and physiological responses measured—say, the amount of vocalisation, and changes in heart rate. We see that Paddy shows a higher level of response on all measured indicators than Sneezy does—her integrated responses score '200', while Sneezy only scores '100'. From this we might conclude that Paddy enjoys her yabbies more than Sneezy does his. However, there is an alternative explanation of the observed data—that Sneezy enjoys his yabbies just as much, or more, than Paddy does, yet he is less reactive and less prone to demonstrate this enjoyment through large changes in the measured indicators.

There is also a possibility that Paddy or Sneezy (or both) actually dislike the yabbies, and their indicators are instead demonstrating the strength of this dislike rather than enjoyment. Perhaps, in a parallel to an 'inverted spectrum' case, they have 'inverted valence', where they display what we would take in ourselves to be markers of positive experience are actually for these animals markers of negative experience. I think we can set this possibility aside, primarily for reasons of parsimony. While it is perhaps logically possible for an animal to display such reactions, in actuality this is highly unlikely, given the probable evolutionary function of such responses in motivating survival-critical behaviours. An animal that, for example, approached the things they hate or fear and avoided the things they love, would be unlikely to succeed. Thus, I will here take it as fixed that the otters have the same indicators of valence—that is, that they will show the same signs of positive and negative experience. We are then only interested in whether the measured levels of response intensity correspond to the same intensity of experience. This is an inference that requires much more justification, as the relationships between the adaptive functions of intensity of feeling and level of response are more likely to be context-specific: for instance, precisely *how* motivated an animal is to obtain a resource will have a greater dependence on individual development and life-history. As it stands, with the data alone we have no way of deciding between the two alternatives I have presented, and thus we cannot make comparisons of welfare. In the next section, I will introduce some methods by which welfare comparisons can proceed.

3 Methods for making welfare comparisons

As I have discussed, welfare comparisons are common within animal welfare science, as well as ethical, political, and institutional decision-making. These decisions require meaningful comparisons of welfare in order to make the requisite inferences and decisions. In this section I'll discuss some of the possible methods for making animal welfare comparisons, starting with a brief overview of the solutions proposed in the human case and why they are largely insufficient. I will describe the potential use of two different similarity assumptions to ground comparisons, and the

---

[2] Yabbies are a type of Australian crayfish.

contexts in which they are most likely to be justified. In Section Four I will address how we might proceed in situations where such justification is not available.

3.1 Solutions in the human case

The problem I have been discussing is in some ways similar to one that is found within the literature on humans—the problem of interpersonal comparisons. This problem has been widely discussed in the literature on human wellbeing, particularly within economics (see e.g. Elster & Roemer, 1991). There are two reasons why I take the problem of interspecies welfare comparisons to be meaningfully different. The first is that most of the discussion of the problem of interpersonal comparisons, and the solutions proposed, rely on a preference-satisfaction view of welfare that is far less convincing for animals, and not the view that is used in this paper. Under a preference-satisfaction account of welfare (an account commonly used for humans) welfare consists in the satisfaction of one's preferences. For animals, while preference satisfaction is often used as an indicator for welfare, to determine which circumstances or resources are likely to improve an animal's welfare, it is far less common as a conception of animal welfare itself (though see Dawkins, 2003, 2017, 2021). Instead, relevance of preferences to welfare can also easily be understood instrumentally within a subjective account of welfare, as representing those things that make an animal feel good or bad; and satisfaction and frustration of preferences themselves producing positive and negative subjective experiences.

An additional reason to think preferences won't be the right account for animal welfare is that individuals can hold preferences for things that are not good for them, such as eating junk food, or taking drugs. In humans, this is usually dealt with through appeal to a modified set of preferences—through constructing some sort of idealised preference set an individual might hold if they had a greater knowledge and understanding of what is good and bad for them. However, this solution appears far less convincing in the case of animals. It is difficult to imagine what an ideally rational version of an otter looks like, or what it would want. By the time we constructed a set of preferences that some ideal otter would hold, taking into account their values, and adding understanding of current and future consequences of actions, this seems to be far from the abilities of any actual otter. If we were able to work according to our ideal otter, this animal would be so vastly different from an actual otter that its preferences may not be a good guide to what our real otters would want, meaning many of their actual preferences would be frustrated. For these reasons, it seems that for animals, preferences are best taken as an instrumental component of subjective welfare, and a good indicator for those things an animal likes or doesn't like, rather than themselves grounding welfare.

The second issue with the solutions that have been proposed within the human case is that they appear insufficient for solving the problem in the animal case. It seems that however bad the problem is in the human case, it's going to be even worse for animals. We might take interspecies welfare comparisons to present a special case of the problem of interpersonal comparisons, but one with enough differences to be worth addressing independently. Firstly, we just don't have as much

information about the minds of animals to work with. In the human case, we can use our knowledge of our own experience and the reported experience of others to make some justified assumptions about similarities and differences between individuals. With animals, as all our information about mental states is coming through indirect measurement of indicators, we cannot be anywhere near as certain. Additionally, making comparisons between members of different species makes the problem even worse, as the differences between individuals will be even larger. Thus, before I move on to discuss what we might do for comparisons of animal welfare, I'll briefly outline the coverage of this problem within the literature on humans and why this does not help for the interspecies case.

There are three main classes of solution proposed in the human case—using an 'introspective' approach to imagine which of two welfare positions is likely to be greater than the other (Binmore, 2009; Harsanyi, 1955), positing a connection between a measurable indicator and subjective experience (List, 2003), and lastly to simply move away altogether from the measurement of subjective welfare and to either measure something else we consider to be important in the questions of distribution under which these comparisons are usually required (such as resource availability), or use a different ethical or distributive principle in decision-making (Fleurbaey & Hammond, 2004). I will briefly examine these here and argue that even if they are potentially useful in the human case (which, for most, is doubtful) they fail to meet the requirements for justifying intersubjective welfare comparisons in animals.

The first potential solution is to use 'imaginative empathy'. This involves an 'introspective' or 'imagining' approach to comparisons, in which an observer assesses the situations of the individuals under comparison and makes an introspective intuitive judgement about their comparative welfare; typically based on imagining themselves in both positions, with each individual's behaviour and desires (and often, forming a judgement about which circumstance they would prefer to be in, taking us back to a preference-based account of welfare). This approach has some deep issues, particularly with reliability. Although we might gain information about the observer's judgement, why should we think that this tracks the fact of the matter about the comparative welfare of the individuals? In the human case, this method is given some (attempted) justification through our understanding of what it is like to be a person under different conditions, with an assumption of similarity between individuals. This relies on our capability of truly imagining ourselves in the place of another, separate from our own desires and psychological biases, and judging between situations. This seems difficult for other humans, and probably impossible when it comes to other species. On what basis could I really make a meaningful judgement about whether a lungfish swimming in a tank is better off than a lion resting on the grass? Both are so far from my own experience that my judgement is certain to reflect mostly my own preferences and biases. It also presupposes that there is some degree of similarity between the experiences of the individuals by which we could make the judgement, and this is partially what is at question. If I don't have information about the intensity of a particular animal's experience, how can I imagine myself in its position at all? In cases where there are disagreements between

observers, there seems no further facts that can be appealed to in order to resolve the dispute, if this is all our comparisons are supposed to rest on.

The second option is to use a 'behaviourist' framework to "posit a fixed connection between certain empirically observable proxies and utility" (List, 2003, p. 1). That is, we use some external indicators of welfare as proxies for the subjective experience, and compare their levels or units instead. In the human case, this relies on using behavioural cues as proxies for internal states. For instance, we might use facial expressions as an indicator of happiness and compare the facial expressions of different individuals as a proxy for comparing their welfare.

This solution begs the question against the very problem we're trying to address. That is, one of the things we are trying to determine is whether (or when) such inferences are justified. It may be the case that they are acceptable in the case of other humans, as we believe our minds to be relevantly similar enough; and we accept verbal reports as sufficient evidence to confirm this. However, it's far less clear that we could take this to be true in other cases involving interspecies comparisons. Why should we believe, for instance, that the same facial expression or vocalisation coming from a lion and from a lungfish represent even the same *type* of mental state, let alone the same intensity of feeling? Now there may be cases in which we are licenced to make this assumption, as I will discuss further in Sect. 3.4, but we cannot take it for granted.

The final option represents essentially giving up on solving the problem of making welfare comparisons, instead moving on to something else; either measurement of an entity that is not welfare (most commonly, money), or a move from scientific questions of measurement to ethical questions of moral status and treatment, such as using a different ethical or distributive principle for decision-making that does not require making comparisons of welfare. Although this may be useful in some cases (as will be described in Section Four), it will not help for cases where we still genuinely wish to make comparisons of welfare, such as those described in Section One. I'll move now to describing the approach I take as being most common within animal welfare science, though it's not made explicit: the use of similarity assumptions.

3.2 Making similarity assumptions to ground welfare comparisons

In order to make animal welfare comparisons in the presence of underdetermining data, as described, we must proceed by making similarity assumptions regarding the presence of (relevant) similarities between different species or individuals. There are two such assumptions that we can make:

Assumption 1 Similarity in capacity.

Assumption 2 Similarity in response.

The first assumption is that the animals have a similar capacity for welfare experience. That is, that the animals are similar in respect to their capacity for overall intensity or scope of welfare experience—the amount of pleasure or suffering they

can and do experience under different conditions. This assumes that the individuals have roughly equivalent minimum and maximum welfare intensities, as well as similar degrees of change in between. The second assumption is that the animals are similar in respect to the level of indicator response shown under the same state of experienced welfare, such as similar heart rate change for mild arousal. By making either of these assumptions, we can essentially overcome the underdetermination problem by holding fixed one source of variation while explaining our data according to the other.

In the ideal case, we can perform welfare comparisons through making both of these assumptions simultaneously. That is, that the animals have both a similar capacity for welfare experience, and a similar response profile. We can test this by mapping the overall response profile for our animals, finding their maximum and minimum response levels for a range of indicators, and the variation across different conditions. In cases where we see animals with a similar profile of indicator responses over different welfare conditions, we can make both assumptions together—that the animals have both a similar scope of welfare experience and a similar degree of response. This is a much more parsimonious explanation than the alternative—that both these factors are varying in tandem in opposite directions to give rise to the seeming similarity. Without a plausible explanation as to why there would be such a hidden difference, we should think that the same responses under the same conditions reflect similarity in the underlying subjective experience, and that our two assumptions hold. The best explanation of observed similarities between the behavioural and other responses of individuals is relevant similarity in underlying subjective states that can ground use of intersubjective comparisons. In this case then, we are justified in making direct comparisons of animal welfare, using the measured indicator responses. However, this circumstance is likely to be rare. More often, we are going to need to make one or the other of the similarity assumptions on their own.

3.3 Assuming similar capacity

The first of these methods of making welfare comparisons thus proceeds through making Assumption 1—that of similar capacity for welfare experience. That is, assuming that the animals are similar regarding their scope of welfare experience (i.e. they have the same scale), but they reflect this differently in their measurable indicator responses. If we make this assumption, we can go on to make comparisons using a zero–one method. In the zero–one method we standardise measures by assigning a score of 0 to the minimum level of welfare and 1 to the maximum level[3] for any individual (Binmore, 2009; Griffin, 1986). Here we assume that the maximum and minimum welfare levels are equivalent between individuals; this is what we're taking for granted under Assumption 1. This provides set points for conversion of individual results onto a common scale.

---

[3] When considering positive and negative welfare we might set these slightly differently—say 1 for best, 0 for neutral and -1 for worst, but the principle remains the same.
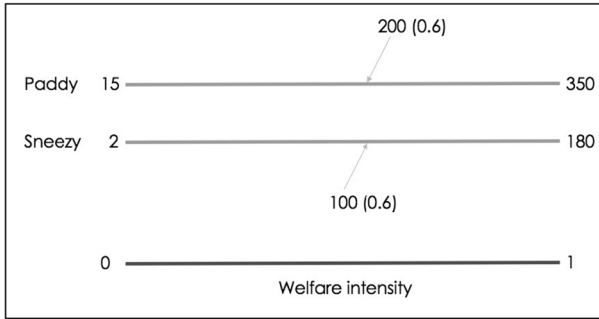
Fig. 1 Comparison of welfare responses under Assumption 1

For each individual, we can build up a welfare profile that measures their level of response under a range of different circumstances to identify where they experience their maximum and minimum welfare levels (0 and 1), and the degree of indicator response they display at these extremes. We can then use these to create a scale for the individual, showing different conditions and indicator responses as proportions of their total, occurring along the 0–1 line. Regardless of the differences between the conditions and indicator responses for individuals, we can still express responses for each as a proportion of the maximum. We then use our assumption about the common value of the 0–1 points to use them as set points to construct a common scale on which comparisons can be carried out.

Consider this method as applied to our otter case. We begin by measuring their individual response profiles. We measure Paddy's responses under different conditions and find that they range from a minimum score of 15 under her most unpleasant condition to a maximum of 350 in her favourite. We then do the same for Sneezy and find a range of 2–180. We then scale these to represent the same range of underlying welfare intensity. On this scale, a score of 350 for Paddy represents the same level of experienced welfare as 180 for Sneezy. So while Paddy might show a response of 200 to yabbies, while Sneezy shows 100—which on the surface makes it seem like Paddy likes them twice as much—when we convert the responses to the 0–1 scale, we find that both are around 0.6 of their maximum response,[4] which would tell us they like them roughly equally (Fig. 1). Paddy is in general more prone to a larger indicator response under all conditions, and so we see that Sneezy's lower absolute response is as high *for him* than Paddy's is for her; we can thus infer that he is actually enjoying the yabbies just as much.

By making the assumption about similarity in degree of welfare (1), we can then use tests under different conditions to measure for differences in the indicator response profiles (2). We hold fixed the level of welfare experience and explain observed variation through differences in the response profiles. Although the

---

[4] I'm here for the sake of simplicity assuming a linear response; this is not necessary for the method, but analysing more complex response curves will obviously require more complex modelling of the data.

example here is an intraspecies one, the method could work equally well for an inter-species case—we can take the range of response scores for our lion and our lungfish and standardise them onto the same scale of welfare intensity in just the same way as described above. All that is required is that our evidence supports this type of similarity between the individuals or species, to justify the use of the assumption.

This assumption appears to be the standard in the small amount of work by welfare scientists writing on the comparison problem in the interspecies case, where it seems to be taken as only a problem of interpreting indicators, taking for granted the similarity in degree of welfare experience across species (Bracke, 2006; Mason, 2010). However, this has so far been done without providing justification. Below, I will examine some of the potential types of justificatory evidence that could be used.

The first line of evidence that could justify use of the assumption is through anal-ogy in the anatomical structures and mechanisms giving rise to welfare experience. Reasoning by analogy holds that where animals are similar in terms of their under-lying structures and mechanisms, they should also be similar in terms of the expe-riences and responses produced. In terms of welfare capacity, similarities in brain structure and function give us reason to think there's similarity in the subjective experience. Brain structure and function will determine the psychology of the indi-vidual, and these will vary depending on the inherited 'instructions' for develop-ment as well as the influence of the developmental environment. Insofar as subjec-tive experience is a function of brain activity, and where there are neural correlates of experience, similarity in brain structure and function (at least within the relevant areas) should then give us similarity in experience.

The level to which we can trust the similarity assumptions will then depend on the level to which there are relevant underlying anatomical and physiological simi-larities. For example, the structures responsible for generating affect appear to be homologous at least across mammals (Berridge & Kringelbach, 2013; Panksepp, 2011), and there are similarities in consciousness-linked brain structures and func-tions at least between vertebrate species if not more widely (Seth et al., 2005). Recent research suggests that despite independent evolutionary events, there are common molecular and neural systems underlying brain organization in different phyla that may be indicative of homology (Strausfeld & Hirth, 2013). The level of similarity will also depend on the degree of variation and plasticity within devel-opmental processes, and further study on the precise mechanisms involved will help determine where the assumptions might hold. Effects of developmental envi-ronment, such as hormonal changes during foetal and infant development, are also likely to play a role in determining capacity for welfare experience. Importantly, our understanding of the mechanisms of sentience will help determine selection of appropriate indicators and how we interpret them. There is thus a strong role here for continuing research into animal sentience, in order to better understand which similarities are relevant.

The second line of evidence is through examination of shared evolutionary his-tory. Animals that have shared evolutionary history, as well as sharing the struc-tures and function of their brains and bodies, also have shared selection pressures. If we take subjective experience, and the behavioural and physiological responses it produces, to be the products of selective processes (e.g. Ginsburg & Jablonka,

2019; Godfrey-Smith, 2017), then it makes sense that shared selection pressures will have led to similar experience and responses. Animals with shared evolutionary history (most particularly those of the same species) will have brains adapted to the same biological challenges, and it makes sense to infer that they will share similar psychology, with the same scope for welfare experience. The minds of different individuals that evolved under the same conditions, are likely to be similar in scope and function. However, how narrowly or widely this justification holds, will depend greatly on the theory of the evolved function of valenced experience. For some, it could hold only at the fine level of similarity found at the species level, while for others it could hold across all sentient life. Again, there is an important role here for continuing research into the evolution of sentience, in order to help settle these questions.

Overall then, it seems that we should use this method—assuming similarity in capacity for welfare experience—in cases where there is evidence for underlying similarities in the structures or processes giving rise to welfare experience, or in its evolved function. There is no simple answer as to when this will be justified—it will depend in large part on the features of the two animals or species being compared, and our understanding of their relevant similarities. It is obviously more likely to hold in cases where the animals are broadly more similar (e.g. same species) than where they are vastly different. Importantly, it will be used in cases where the evidence suggests that this type of similarity is more likely than similarity in response profiles. In other cases, we should instead take the alternative method based on making Assumption 2.

3.4 Assuming similar response

The alternative to the above method of making welfare comparisons is through making Assumption 2—that of similar response. Here, we take it to be the case that the animals' response indicators reflect the same underlying welfare experience—that a score of 10 for one animal represents the same absolute amount of welfare as a 10 for another animal. By taking this assumption, we're then able to use behavioural and physiological data to determine the intensity of their subjective experience under different conditions (as well as to map out the maximum and minimum overall levels).

We can see how this works by turning again to the case of our otters. We once more map out their range of responses under different conditions, finding Paddy varies from 15 to 350 and Sneezy from 2 to 180. But this time, instead of scaling the responses onto the same underlying scale, we instead compare the absolute responses. As we're assuming similarity in the response relationship, we take these scores as directly representative of experienced welfare, in the same way for each otter. Paddy's higher reaction levels suggests she is capable of experiencing more pleasure than Sneezy under a range of circumstances. Her 'highs' are higher, while his 'lows' are lower.

Comparing again their reactions on receiving the yabbies—Paddy showing a score of 200 to Sneezy's 100—we then take Paddy's more extreme reactions to
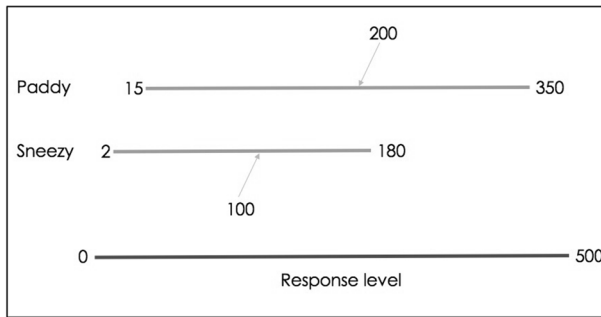
Fig. 2 Comparison of welfare responses under Assumption 2

mean that she is indeed experiencing more pleasure (twice as much) in receiving the yabbies (Fig. 2). By making the assumption about similarity in indicator response (2), we're able to run tests to measure the differences in welfare intensity experienced by individuals. We hold fixed the relationship between welfare experience and indicator response and explain observed variation through differences in the underlying experience of welfare. Using Assumption 2 justifies drawing the conclusion that Paddy enjoys her yabbies more than Sneezy does.

In this case, by making the assumption about similarity in indicator response level (2), we can then use our tests to make direct inferences about the underlying differences in welfare (1). We hold fixed the indicator response profile to explain observed variation through differences in the intensity of welfare experience. Again, this method could equally well be applied in an interspecies case—we examine the range of response indicators for our lion and our lungfish and compare their absolute values as representative of underlying welfare experience. The method can be used whenever our evidence supports this type of similarity between the individual animals or species, to justify use of the assumption.

Our first line of evidence will be analogy in the structures or processes that give rise to the particular indicator responses we are using. For instance, where neural structures directly mediate indicator responses, similarity in neural systems will also give us reason to think there will be similarity in these responses. Often, though, indicator response will also involve other physiological pathways and in these cases, we would also require similarities in the relevant response-producing mechanisms—such as the hormonal and neuronal outputs of the brain, and their impacts on bodily systems—to justify concluding there's similarity in the responses produced. This will rely on the indicators we're using, and the proposed mechanism for linking these indicators to welfare. For example, for indicators with a more flexible developmental pathway (e.g. behaviour) we would be more inclined to assume that response levels vary quite a lot between species. There are many examples of these differences—urination and defecation in a new environment is a scent-marking behaviour in mice but a sign of fear in rats, and bulls show decreased corticosteroid response after tethering, while pigs show increased response (Mason & Mendl, 1993). For more deeply physiologically controlled indicators (e.g. heart rate) with pathways

that appear to vary less between individuals, we would be more likely to assume that responses reflect experience more directly, where response profiles are likely to be similar. Here, we require evidence from a range of biological sciences relevant to the indicator responses of interest, such as animal welfare science, animal behaviour, and physiology.

The second justification is that of evolutionary history. Physiological and behavioural responses to subjective welfare changes are going to depend in large part on evolutionary history, and in many cases we might see similar selection processes operating on the indicator responses. Again, this will depend very much on the specific indicator used and its function in different animals. For behavioural responses this will include what was beneficial to communicate with others—for example, prey species are notoriously non-vocal when in pain as they do not wish to alert predators to their weakened status. These responses are much more likely to be narrowly similar, perhaps only among conspecifics. For physiological responses this is likely to include those responses appropriate to ready the body to meet whatever particular challenges it is about to face, such as activation of the HPA axis under stress. These responses could be much more broadly preserved across diverse taxa.

Finally, another justification for use of this assumption will occur when we see convergence between difference indicators, using a form of robustness reasoning (Wimsatt, 2007). A phenomenon is robust if it's observed across a range of different tests, each of which rely on different background assumptions. By testing the response profile for an individual across a range of indicators, we can get a better idea as to whether observed variance is likely to be a result of variance in underlying welfare state or in indicator response. If the different indicators give us similar results (e.g. one individual shows higher overall response across a range of indicator types), then this gives us reason to think that this is reflective of differences in underlying welfare intensity. If instead we see different results across indicator types, this gives us reason to think that it's the indicator response profiles that are varying, and it is more likely that Assumption 1 holds and welfare intensity is similar.

This relies on the assumption that different indicator types really are produced through relevantly independent mechanisms, which can only be supported as we know more about the mechanisms of welfare experience and response production. If, for example, all indicators are 'centrally' controlled through a single initial effect, such as signal output from one brain region in response to a welfare change, they will not really count as independent for the purposes of testing. The good news is that this is itself a testable prediction—both through examining the pathways through which responses are produced, and by looking for degree of correlation of indicators under different conditions.

There is one method that is sometimes used for making welfare comparisons that I take as a special version of Assumption 2, and this is the use of 'sentience proxies'. Often, the question of interspecies comparisons is framed as a question about the comparative level of sentience of different organisms (i.e. their capacity for a particular scope of intensity of valenced experience). This has been called the 'emotional capacities claim' (Višak, 2017); that animals with stronger emotional capacities are capable of higher welfare states. This is a natural consequence of accepting a subjective experience account of welfare, as I have here. Here then, one can try to

make comparisons through use of alternative proxies for sentience or welfare capacity, typically some form of neural or cognitive complexity (e.g. Budolfson & Spears, 2020). There are different possible proxies, including brain size, number of neurons, and connective complexity. I won't assess the quality of these different suggestions here, but will point out that their use relies on a background assumption of similar response profile. That is, that different species show the same relationship between the proxy and welfare capacity, such as an increase in capacity with increasing neural complexity.

This again is an assumption that requires justification—strong justification in the case that it is used across a wide range of taxa—and one that will rely on better developing our science of animal sentience to confirm the substrates and processes that generate conscious experience and how they scale with intensity of experience. Some headway could be made with looking for correlates of different intensity of experience, as established through human self-report paradigms showing correlation between subjective report of intensity of experience, behavioural responses, and intensity of brain activity (Coghill et al., 2003); but these must be taken with caution without embedding within a richer framework of understanding sentient experience. Though there has been some promising work on identifying the brain regions and level of activity associated with positive and negative valence (e.g. Berridge & Kringelbach, 2013; Davidson, 1992; Panksepp, 2011). we are still a long way from a reliable neuroscience of affect, particularly one strong enough to ground interspecies comparisons. Still, though we may still have a long way to go, the neuroscience of affect is growing all the time (for a recent review see Paul et al., 2020) and further research into the neurobiology of affect intensity might thus be considered a high priority for answering this question.

The method of assuming similarity in response profile is best used in cases where there is evidence for underlying similarities in the structures or processes that generate the indicator responses we are measuring. This will again be dependent on the traits of the species of interest, and their relevant similarities. It is also important to determine whether the evidence suggests that this type of similarity is more likely than similarity in capacity for welfare experience, or whether we should be making Assumption 1.

I have here outlined three methods for making welfare comparisons, based in accepting one, or both, of the similarity assumptions I have described. Our confidence in the comparisons will be related to the strength of justification for the similarity assumptions, which will vary depending on contextual features of the animals we are comparing, the types of similarities they possess, and our understanding of which are the most relevant similarities. There is a highly important role here for further research to understand the relevant processes giving rise to welfare experience and producing indicator responses, and thus to provide justifications for the use of these assumptions. In all cases, we should choose the option that has the stronger justification. In particular, we can allow for varying degrees of similarity with differing strengths of justification; and as I will discuss further in Section Five, our willingness to accept these will depend on the context and degree of confidence required for our purposes. However, in some (perhaps many) cases, we may not have strong justification for either.

The similarity assumptions largely rely on justifications of analogy and shared evolutionary history that will only hold for animals that share the relevant similarities in physiology or evolutionary history—typically those of the same species, or perhaps closely related species. This might also require those with similar developmental histories, which may mean for example sub-groups within species of age, sex, or rearing type (wild vs captive). There are known effects of individual personality and temperament—as well as genetics and early experiences—on emotional responses to stimuli, and thus welfare (Boissy et al., 2007). Again, we need to understand how particular anatomical structures and biological processes give rise to both the subjective experience of welfare, and the indicators that we use to measure it, in order to identify the relevant similarities for welfare comparisons and which groups of animals possess them. Understanding the extent of similarity in structure, function and selection pressures across different groups will help us see how far we might extend this solution. For example, if we found that the mechanism linking welfare experience to changes in heart rate was one which arose fairly early in evolution, shared across all vertebrates, then we could use this indicator to make comparisons between animals within this entire group.

However, while these might provide good solutions in most of the comparisons made between members of the same species, this is far less likely to be the case for interspecies comparisons. We do not have good reason to think that the similarities hold between quite dissimilar species. The types of indicators tend to be quite species-specific (especially behavioural indicators), and we should be circumspect in inferring similarity between species. Think again of our zoo manager trying to compare lions and lungfish; two species so disparate that it's unlikely that the similarity assumptions will apply. It's perfectly plausible that lungfish and lions have completely different scopes for intensity of welfare; so that the heights and depths of lion experience may just be of a different scale to that of lungfish. It's also extremely improbable that there will be overlap in the types of indicators used to measure welfare in each species, let alone that they will be subject to the same processes linking subjective welfare to indicator outcomes. There does not seem to be any objective standard to which we can appeal in order to convert units of lion welfare into units of lungfish welfare, and so we cannot make meaningful comparisons. But we still want to have some means of comparing the welfare gain to the lion from its underfloor heating to the gain of the lungfish of having new logs to explore and shelter in. What, then should we do in these cases?

4 Alternative methods

We've seen that there are several possible methods for making welfare comparisons, through the use of similarity assumptions that hold fixed one potential source of variation to explain observed data in terms of the other; allowing us to draw conclusions about relative welfare. In cases where the relevant similarities hold—whatever they might end up being—we're able to make the required assumptions and so perform intersubjective comparisons of welfare. However, these methods are limited to the contexts in which we can justify sufficient level of similarity along one of

the two dimensions—similarity in welfare capacity, or similarity in response profile. This means that while they are likely to hold in many cases of within-species comparisons, they are unlikely to be of use in many of the important interspecies contexts discussed in Section One. We can hope that continued research into the biological and evolutionary basis of subjective experience may in future provide sufficient grounding for a wider range of comparisons. However, in the meantime we still may be required to act in these contexts, and need some way of proceeding. In this section I'll describe some possible alternative methods we may use in the cases where similarity assumptions fail, discussing two options—use of moral weights, and recourse to other ethical or distributive principles. The obvious problem with these methods is that they are not direct solutions to the empirical problem of welfare comparisons, and as such may be unsatisfactory in many cases. However, they may still serve as pragmatic guides for action in the contexts in which comparisons are required but cannot yet be adequately justified by the similarity assumptions described above.

4.1 Moral weight

The first option is to essentially shift from the empirical problem to the moral problem and switch to a comparison of moral value, rather than directly of sentience or welfare experience. This way, we are simply comparing animals based on how much they matter within our ethical framework. In some cases, moral value could be a function of welfare capacity. However, where (as described) we do not have enough information about comparative capacity, we could then use another method of determining comparative moral status. This could be set via convention—such as a consensus on how much we want the interests of a particular animal or species to count in our deliberations—or via some proxy (including the sentience proxies described earlier).

One such example is an equal consideration of individuals, where the welfare of each individual is given the same weighting, regardless of absolute strength; a *presumption* of equal status (Zuolo, 2017). This would ensure that a lungfish gets its best possible welfare and a lion gets its, despite potential differences in intensity between them. That is, we could say that allowing a lungfish to achieve its maximum welfare level is of equal moral importance as allowing a lion to achieve its, even if it turns out that the lion actually experiences three times the welfare intensity at its maximum than the lungfish do.

Once a moral weighting is set, this can be used as a way of scaling animal interests within cost–benefit or resource-distribution calculations such as were described in the beginning of the paper. However, there is a risk here of arbitrariness. Whatever principles we use for ethical decision-making in these cases of uncertainty about comparative welfare, they are likely to be specific to context and background values, such as how much importance we place on equality, or suffering. For those (such as utilitarians) strongly committed to ensuring that equal experiences count equally, a different method of setting moral weight may come up with entirely the

wrong results. However, in the absence of determinate information about relative sentience or welfare experience, it might be the best we can do.

4.2 Other decision-making procedures

Another method is to move to an alternative ethical framework or distributive principle that does not require direct welfare comparisons. We can try to switch from a utilitarian framework of welfare maximisation to another distributive principle to allow decision-making without comparisons. Many other principles will not be suitable as they also require comparisons between individuals. This includes prioritarian or egalitarian distributive rules that prioritise improving the situation of the worst-off, or ensuring equal distribution between all individuals; these will still require us to make comparisons to identify relative welfare levels. The preferred method for human cases is use of the Pareto rule—ensuring that all either improve their situation or end up no worse off (Fleurbaey, 2016)—that only requires intrasubjective level comparisons to assess whether individuals are being made better off. While this is an intuitively appealing principle, it is of limited use in most situations, as resource scarcity and competing interests make it impossible to improve the welfare of all individuals. At times it may be the case that we have to accept a decrease in welfare of one individual or group to create a larger increase in welfare for another group. It also remains silent on how to choose between options in which there is a benefit for one group rather than another, even if none are made worse off (think again of our lion and lungfish).

While there are clear limitations to these methods, there will be many contexts in which they will be sufficient for the task. Despite how one decides to make decisions in these cases, what is most important to highlight is that we shouldn't attempt to make direct comparisons of welfare in cases where this is not justified, as this is highly unlikely to lead to reliable results of the type we want. Our confidence in judgements drawn from the methods of comparison that rely on similarity assumptions will only be as strong as the justification for those assumptions and, as discussed, sometimes this will be quite weak indeed. In contexts where comparisons are being made for purposes for which an alternative decision-making procedure may suffice, these options should be considered.

5 Conclusion

One of our biggest problems in animal welfare science and its applications is our ability to make welfare comparisons between different individuals, particularly different species. In this paper, I've outlined the scope of the problem and assessed the available methods without proposing a single specific solution; there's no 'magic bullet' to solve this complex problem. This work should be seen as an attempt to map out the space of possible methods for performing welfare comparisons, and identifying when and how their use might be justified, including methods by which it might be strengthened and performed more systematically. In this sense, it's a strongly

pragmatic discussion, framed around the idea that we do make such comparisons, and we want to know how we can do them better within our current constraints.

The empirical problem of welfare comparisons—our ability to determine the appropriate 'conversion formula' between different welfare scales—rests on an underdetermination problem where our data does not uniquely determine our conclusions but can instead support different interpretations given different plausible background assumptions. We cannot distinguish between the two sources of variation that can explain our results—variation in the underlying target variable (welfare experience) and in the relationship of measured indicators to the target. When welfare scientists make comparisons, they are typically making implicit similarity assumptions that hold fixed one of these sources of variation to explain the data in terms of the other. These assumptions can be justified by analogy and shared evolutionary history, however will only hold in cases where individuals possess the relevant underlying similarities. In cases of comparisons across species, we will often not be able to justify such assumptions and instead may need to use different ethical or distributive principles to make the decisions for which we would otherwise want to use comparisons.

A final possible alternative is to abandon the subjective conception of welfare in favour of another account; perhaps even with the argument that failure to solve the problem of interspecies comparisons should count against the use of this concept in the first place. How, then, would the other primary concepts of animal welfare cope with this problem?

As mentioned previously, the three main alternative welfare concepts are the biological functioning account, the natural living account, and the preference-satisfaction account. Biological functioning takes animal welfare to consist in good physical functioning—the health and fitness of the animal. This may prove extremely difficult to operationalise in practice for the purposes of comparisons. While we can measure a range of physical parameters, there is no obvious privileged set of weightings with which to combine them into a single 'health' score; and as such we would be left with a new set of problems to address, such as: is a lion with a kidney infection in a better or worse state than a lungfish with a fin injury? It is not clear that we would be in any better a position than with the subjective welfare concept in addressing the interspecies comparisons problem.

Natural living takes animal welfare to consist in the animal living according to its evolved functions, sometimes called *telos*, usually with a particular focus on performance of natural behaviour. There are deep methodological issues in attempting such a project (see Veasey et al., 1996a, 1996b), but perhaps we could imagine at least in theory assigning each animal a 'naturalness' score and comparing these to determine which animal has the better welfare. The main problem with this method is that it leads to highly counterintuitive results—for example, that humans, with our highly 'unnatural' lifestyles, will have very low welfare scores, while any wild animal, even as it is being eaten by a predator or dying of thirst during a drought, would have a high score. Attempts to refine the naturalness concept to rule out such cases, such that it only emphasises some relevant subset of natural states/behaviours will then typically require a filtering criterion such as effect on subjective experience, which then leads us back to this welfare concept (Browning, 2020b).

Finally, there is preference-satisfaction, that takes animal welfare to consist in an animal having its preferences satisfied. While, as discussed in Sect. 3.1, this account of welfare is often adopted in the human case to try and circumvent the problem of interpersonal comparisons, it is unlikely to be successful here. Firstly, as already mentioned, it is not a particularly plausible account of animal welfare. Secondly, even if it were adopted, it would not solve the problem. Instead, comparing preferences directly runs into the same problem of commensurability for strength of preferences—is the most preferred option for a lion preferred just as strongly as the most preferred option for a lungfish, or can one species simply hold stronger desires than another? Unless one adopts the similarity assumptions I have described—in particular, a version of the 'similar capacity' assumption that allows us to use a zero–one rule to standardise the strength of the most and least preferred options across individuals—then the same problem will remain.

In the end, none of the available methods for making welfare comparisons are ideal. They rely on background assumptions for which there's currently limited empirical support, and which are likely to apply in a limited range of cases; or they require giving up on the project of trying to make welfare comparisons at all. Which method we choose is thus going to be highly context-specific. Decisions should be made with an honest assessment of the purposes for which we require the comparison and which method may be best for the task, while acknowledging the potential limitations and drawbacks.

It's not necessary that our methods are perfect. It's important to keep in mind that comparisons are made for a reason, and we only need to be confident enough in our comparisons to serve the reason at hand. Our level of confidence in the method thus only needs to match what is required for the application. For most practical purposes, certainty is not required; just a reasonable assumption that we're getting close to the fact of the matter about comparative welfare experience. It may not always be the case that a lot hangs on getting it exactly right. In many applications, such as deciding on resource distribution, we can be confident that we have made a welfare improvement, even if it were to turn out that this was not the maximally efficient use of our resources to do so.

What is important is that we make our choices, and our assumptions, explicit. This allows transparency in the process, as well as correction when further empirical data comes to light. More than anything else, we're in need of further research. Only through gaining a better understanding of the subjective experiences that make up welfare—how they evolved and how they function—can we establish a firm empirical base from which to justify our comparisons.

References

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

Binmore, K. (2009). Interpersonal comparison of utility. In H. Kincaid & D. Ross (Eds.), *The Oxford handbook of philosophy of economics* (pp. 540–559). Oxford University Press.

Boccia, M. L., Laudenslager, M. L., & Reite, M. L. (1995). Individual differences in macaques' responses to stressors based on social and physiological factors: Implications for primate welfare and research outcomes. *Laboratory Animals, 29*(3), 250–257. https://doi.org/10.1258/002367795781088315

Boissy, A., Manteuffel, G., Jensen, M. B., Moe, R. O., Spruijt, B., Keeling, L. J., Winckler, C., et al. (2007). Assessment of positive emotions in animals to improve their welfare. *Physiology & Behavior, 92*(3), 375–397.

Bracke, M. B. M. (2006). Providing cross-species comparisons of animal welfare with a scientific basis. *NJAS - Wageningen Journal of Life Sciences, 54*(1), 61–75. https://doi.org/10.1016/S1573-5214(06)80004-7

Browning, H. (2020a). *If I Could Talk to the Animals: Measuring Subjective Animal Welfare*. PhD Thesis, Australian National University. https://openresearch-repository.anu.edu.au/handle/1885/206204

Browning, H. (2020b). The natural behavior debate: Two conceptions of animal welfare. *Journal of Applied Animal Welfare Science, 23*(3), 325–337. https://doi.org/10.1080/10888705.2019.1672552

Browning, H. (2022a). Assessing measures of animal welfare. *Biology and Philosophy, 37*(36), 1–24. https://doi.org/10.1007/s10539-022-09862-1

Browning, H. (2022b). The measurability of subjective animal welfare. *Journal of Consciousness Studies, 29*(3–4), 150–179. https://doi.org/10.53765/20512201.29.3.150

Budolfson, M., & Spears, D. (2020). Quantifying animal well-being and overcoming the challenge of interspecies comparisons. In B. Fischer (Ed.), *The Routledge handbook of animal ethics* (pp. 92–101). Routledge.

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

Davidson, R. J. (1992). Emotion and affective style: Hemispheric substrates. *Psychological Science, 3*(1), 39–43. https://doi.org/10.1111/j.1467-9280.1992.tb00254.x

Dawkins, M. S. (2003). Behaviour as a tool in the assessment of animal welfare. *Zoology, 106*(4), 383–387.

Dawkins, M. S. (2017). Animal welfare with and without consciousness. *Journal of Zoology, 301*(1), 1–10. https://doi.org/10.1111/jzo.12434

Dawkins, M. S. (2021). *The science of animal welfare: Understanding what animals want*. Oxford University Press.

Duncan, I. J. (2002). Poultry welfare: Science or subjectivity? *British Poultry Science, 43*(5), 643–652.

Elster, J., & Roemer, J. E. (1991). *Interpersonal comparisons of well-being*. Cambridge University Press.

Fleurbaey, M. (2016). Economics and economic justice. In Zalta, E. N. (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition). Available at: https://plato.stanford.edu/archives/win2016/entries/economicjustice/

Fleurbaey, M., & Hammond, P. J. (2004). Interpersonally comparable utility. In S. Barberà, P. J. Hammond, & C. Seidl (Eds.), *Handbook of utility theory: Volume 2 extensions* (pp. 1179–1285). Springer, US.

Fraser, D., Weary, D. M., Pajor, E. A., & Milligan, B. N. (1997). A scientific conception of animal welfare that reflects ethical concerns. *Animal Welfare, 6*(3), 187–205.

Ginsburg, S., & Jablonka, E. (2019). *The evolution of the sensitive soul: Learning and the origins of consciousness*. MIT Press.

Godfrey-Smith, P. (2017). Animal evolution and the origins of experience. In D. L. Smith (Ed.), *How biology shapes philosophy* (pp. 51–71). Cambridge University Press.

Griffin, J. (1986). *Well-being: Its meaning, measurement and moral importance*. Clarendon Press.

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy, 63*(4), 309–321.

Izzo, G. N., Bashaw, M. J., & Campbell, J. B. (2011). Enrichment and individual differences affect welfare indicators in squirrel monkeys (*Saimiri sciureus*). *Journal of Comparative Psychology, 125*(3), 347–352.

List, C. (2003). Are interpersonal comparisons of utility indeterminate? *Erkenntnis, 58*(2), 229–260.

Manteca, X., & Deag, J. M. (1994). Individual variation in response to stressors in farm animals: Implications for experimenters. *Animal Welfare, 3*(3), 213–218.

Mason, G. (2010). Species differences in responses to captivity: Stress, welfare and the comparative method. *Trends in Ecology & Evolution, 25*(12), 713–721. https://doi.org/10.1016/j.tree.2010.08.011

Mason, G., & Mendl, M. (1993). Why is there no simple way of measuring animal welfare? *Animal Welfare, 2*(4), 301–319.

Mellor, D. J. (2016). Updating animal welfare thinking: Moving beyond the 'five freedoms' towards 'a life worth living.' *Animals, 6*(3), 21.

Nielsen, C. S., Price, D. D., Vassend, O., Stubhaug, A., & Harris, J. R. (2005). Characterizing individual differences in heat-pain sensitivity. *Pain, 119*(1–3), 65–74. https://doi.org/10.1016/j.pain.2005.09.018

Panksepp, J. (2011). Cross-species affective neuroscience decoding of the primal affective experiences of humans and related animals. *PLoS ONE, 6*(9), e21236. https://doi.org/10.1371/journal.pone.0021236

Paul, E. S., Sher, S., Tamietto, M., Winkielman, P., & Mendl, M. T. (2020). Towards a comparative science of emotion: Affect and consciousness in humans and animals. *Neuroscience & Biobehavioral Reviews, 108*(January), 749–770. https://doi.org/10.1016/j.neubiorev.2019.11.014

Pettigrew, R. (2019). *Choosing for changing selves* (1st ed.). Oxford University Press. https://doi.org/10.1093/oso/9780198814962.001.0001

Seth, A. K., Baars, B. J., & Edelman, D. B. (2005). Criteria for consciousness in humans and other mammals. *Consciousness and Cognition, 14*(1), 119–139. https://doi.org/10.1016/j.concog.2004.08.006

Singer, P. (2016). Afterword. In T. Višak & R. Garner (Eds.), *The ethics of killing animals* (pp. 229–236). Oxford University Press.

Strausfeld, N. J., & Hirth, F. (2013). Deep homology of arthropod central complex and vertebrate basal ganglia. *Science, 340*(6129), 157–161. https://doi.org/10.1126/science.1231828

Veasey, J. S., Waran, N. K., & Young, R. J. (1996a). On comparing the behaviour of zoo housed animals with wild conspecifics as a welfare indicator. *Animal Welfare, 5*, 13–24.

Veasey, J. S., Waran, N. K., & Young, R. J. (1996b). On comparing the behaviour of zoo housed animals with wild conspecifics as a welfare indicator, using the giraffe (*Giraffa camelopardalis*) as a model. *Animal Welfare, 5*, 139–153.

Višak, T. (2017). Cross-species comparisons of welfare. In A. Woodhall & G. G. da Trindade (Eds.), *Ethical and political approaches to nonhuman animal issues* (pp. 347–363). Springer International Publishing. https://doi.org/10.1007/978-3-319-54549-3_15

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Harvard University Press.

Zuolo, F. (2017). Equality, its basis and moral status: Challenging the principle of equal consideration of interests. *International Journal of Philosophical Studies, 25*(2), 170–188. https://doi.org/10.1080/09672559.2017.1286679