The problem of interspecies welfare comparisons (preprint)

Heather Browning

(ORCID 0000-0003-1554-7052)

drheatherbrowning@gmail.com

London School of Economics and Political Science, Centre for Philosophy of Natural and Social Science
Houghton Street, London, WC2A 2AE, UK

Abstract

One of the biggest problems in applications of animal welfare science is our ability to make comparisons between different individuals, particularly different species. Although welfare science provides methods for measuring the welfare of individual animals, there's no established method for comparing measures between individuals. This problem occurs because of the underdetermination of the 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. In this paper I describe the similarity assumptions that underlie our current applications of interspecies comparisons and examine in which cases they are justified, as well as describing alternative methods we may use when they are not. In the end, all our available options for making interspecies 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 welfare comparisons.

Keywords: animal welfare, comparison, measurement, underdetermination, interspecies

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 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 and Spears 2020). 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. 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.

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, as well as briefly discussing some proposed solutions to the comparisons problem in the human case, and why they fail. In Section Three I'll introduce the background assumptions that are used to make welfare comparisons, illustrating with a worked example. In Section Four I'll discuss in which cases this approach is justified and in Section Five address what we can do in the cases where it is not, before moving on in Section Six 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. Given that, as I have discussed, we do frequently perform such comparisons, the problem is one of justifying our methods, or finding the best available. 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 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 the absence of a solution, I'll engage a little of the ethical theory.

In large part, the question of interspecies comparisons is a question about the comparative level of consciousness, or sentience, of different organisms. 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 hedonic or subjective experience account of welfare, as I will for this paper. This is a common view within animal welfare science (e.g. Duncan 2002; Mellor 2016), and one of the primary views within animal ethics (e.g. Singer 2016). Throughout this paper I'll use 'welfare' to refer simply to this aspect: the integrated set of valenced mental states or 'affects'. 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.

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

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

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 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.

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. 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. This is not just hypothetical – within-species differences in individual behavioural and physiological responses to positive and negative stimuli are common (e.g. Boccia, Laudenslager, and Reite 1995; Izzo, Bashaw, and Campbell 2011; Manteca and Deag 1994), and it's difficult in these cases to determine whether or not results imply a welfare difference. 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. Without such information, we cannot make comparisons.

Before I move on to discuss what we might do, I'll briefly touch on the coverage of this problem within the literature on humans and why this does not help for the interspecies case. The problem of making interpersonal comparisons of welfare has been widely discussed in the literature on human wellbeing, particularly within economics (see e.g. Elster and Roemer 1991), though often only with a preference-satisfaction view of welfare in mind. It seems that however bad the problem is in the human case, it's going to be even worse in the animal case. 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, we will often want to make comparisons between members of different species, and this will make the problem even worse, as the differences between individuals will be even larger.

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 and Hammond 2004). I'll not detail these possibilities here, but I do not think they are sufficient to deal with the problem of making intersubjective welfare comparisons for animals. The first approach relies on a reliability of judgement about the internal states of others that, even if justified in the case of humans, seems almost certainly useless in the case of other species. The second approach is equivalent to what is already being done in welfare science, but as I will discuss in the next section, implicitly relies on background assumptions that may not be justified in many cases. 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. Although this may be useful in some cases (as will be described in Section Five), it will not help for cases where we still genuinely wish to make comparisons of welfare. 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. The use of similarity assumptions in 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 look at how these comparisons are usually performed, and the implicit background assumptions that underlie them. As these are typically done intuitively and/or unreflectively, without consideration for the problems discussed, it's primarily important that those performing comparisons are able to identify and make explicit the background assumptions they are relying on.

Comparisons of animal welfare rely on 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 experience

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 level of welfare intensity - 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. I will shortly move on to demonstrate how this works using an example, but first I want to examine another question – are these assumptions justified?

3.1 Justifying similarity assumptions

There are two primary lines of reasoning and evidence that can be used to support the similarity assumptions – analogy and shared evolutionary history. Reasoning by analogy holds that where animals are similar in terms of their underlying structures and mechanisms, they should also be similar in terms of the experiences and responses produced. Many animals have such similarities in their anatomy and physiology; the structures and mechanisms that give rise to both subjective experience and indicator responses.

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 individual, and these will vary depending on the inherited 'instructions' for development as well as the influence of the developmental environment. Insofar as subjective experience is a function of brain activity, and where there are neural correlates of experience, similarity in brain structure and function should then give us similarity in experience. 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 give us reason to think there's similarity in the responses produced.

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 similarities. For example, the structures responsible for generating affect appear to be homologous at least across mammals (Berridge and Kringelbach 2013; Panksepp 2011), and there are similarities in consciousness-linked brain structures and functions at least between vertebrate species if not more widely (Seth, Baars, and Edelman 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 and Hirth 2013). The level of similarity will also depend on the degree of variation and plasticity within developmental processes, and further study on the precise mechanisms involved will help determine where the assumptions might hold. Effects of developmental environment, such as hormonal changes during foetal and infant development, and conditioning of behavioural responses throughout life, are also likely to play a strong role in determining both scope of experience and level of responsiveness. Importantly, our understanding of the mechanisms of sentience will play a large role in determining selection of appropriate indicators and how we interpret them.

The second justification is that of evolutionary history. Animals which have shared evolutionary history, as well as sharing the structures 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, as being the products of selective processes (e.g. Ginsburg and 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. Physiological and behavioural responses to subjective welfare changes are going to depend in large part on evolutionary history. 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. 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 increased heart rate for fight or flight. The minds and the indicators of different individuals that evolved under the same conditions, are likely to be similar in scope and function.

Finally, these can be strengthened through considerations of parsimony. This is using an inference to the best explanation; one that can describe our results within our best frameworks of understanding. This would include not positing differences without evidence of their existence – although it may be *possible* that different individuals vary widely in experience and response, if there's no positive evidence of this fact, then it does not seem likely that this is the case. Overall, it seems far more likely that individuals with these similarities in anatomy, physiology and evolutionary/developmental history will have broadly similar minds than vastly different ones. Without evidence to the contrary, it seems more parsimonious to assume similarity in these cases.

If the minds are sufficiently similar, then we're justified in making our assumptions and the comparisons which rely on them. In particular, we can allow for varying degrees of similarity with differing strengths of justification; and as I will discuss further in Section Six, our willingness to accept these will depend on the context and degree of confidence required for our purposes. Importantly, as mentioned, use of these assumptions allow us to hold fixed one source of variation and make inferences or draw conclusions based on the other, using this to make meaningful intersubjective welfare comparisons.

3.2 Using similarity assumptions in practice

This method - using similarity assumptions to ground conclusions from comparative welfare data - is how I take animal welfare science to proceed in practice, though without making the process explicit. Here, I'll provide an example illustrating how this can occur. Note that this is an example of an *intra*species comparison, used for the sake of simplicity and because the similarity assumptions are most easily justified in these cases, but should help illustrate the additional difficulties that occur for interspecies cases.

I used to work with two otters – Sneezy and Paddy. Imagine that each are given some yabbies, 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.

How do we arrive at this conclusion? Through making our second assumption – that of similar response. That is, 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 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). Let's see how this works in the case of our otters.

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. 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 mean that she is indeed experiencing more pleasure (twice as much) in receiving the yabbies (Figure 1). 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. I take this to be the most common way welfare comparisons are performed in practice, at least within species.

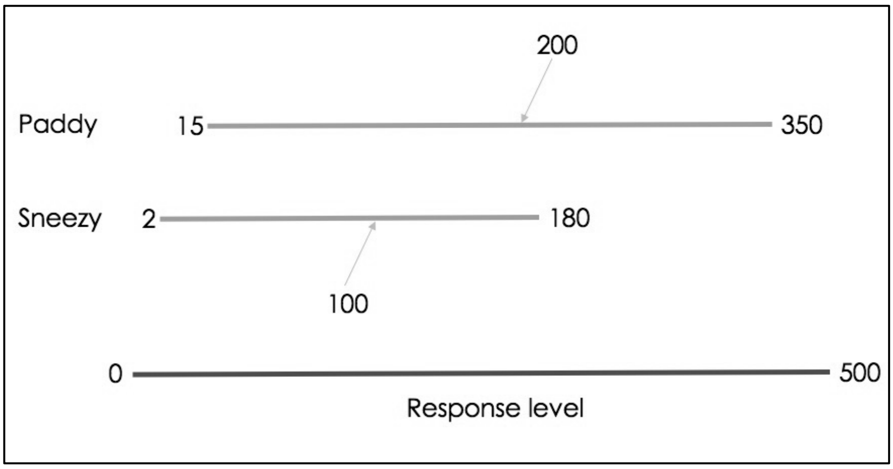

Figure 1: Comparison of welfare responses under Assumption 2

Using Assumption 2 justifies drawing the conclusion that Paddy enjoys her yabbies more than Sneezy does. However, there's another conclusion we could draw from the same data: that Sneezy enjoys his yabbies just as much (or even more) than Paddy does, but he is much less reactive. This conclusion could be justified through appeal to the first assumption – that of similar experience. That is, that both 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. 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 (Bracke 2006; Mason 2010).

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[2] 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.

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

[2] 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.

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 construct a common scale on which comparisons can be carried out.

We again map out their range of responses under different conditions, finding Paddy varies from 15 - 350 and Sneezy from 2 – 180. But this time, instead of comparing the absolute responses, we scale these to represent the same range of underlying welfare levels. 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 scale to the 0-1 scale, we find that both are around 0.6 of their maximum response[3], which would tell us they like them roughly equally (Figure 2). 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 degree of welfare intensity and explain observed variation through differences in the response profiles.

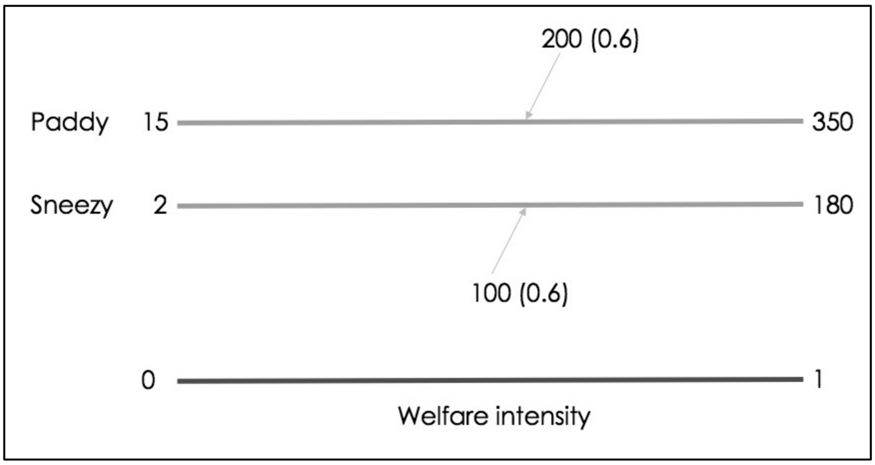

Figure 2: Comparison of welfare responses under Assumption 1

[3] 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.

So we have two ways of making these comparisons, based on making one of two different similarity assumptions – assumption of similar experience, or assumption of similar response. Once we make one of these assumptions, we can then use our data to draw conclusions about the comparative welfare of the different animals. The problem then remains – how do we determine which assumption (if either) will be relevant in different cases? And what can we do in cases where this approach fails?

4. Determining which assumption to use

Although the method of making similarity assumptions to ground welfare comparisons is common in animal welfare science, as we have seen there are two such assumptions that can be made, each of which will lead to a different interpretation of our data and drawing of different conclusions. The data itself underdetermines which result we should prefer, with no obvious and unproblematic way to choose between them. The problem is then to identify which assumption we should make and to justify its use. There are three possibilities I will explore in this section – the best-case scenario of making both assumptions together, the intermediate case of making one assumption or the other, and the worst case, in which we're not able to make either assumption as we have both types of variance occurring at once.

4.1 Making both assumptions

In the best possible case, we're able to make both assumptions together. That is, that the animals have both a similar welfare experience, and a similar response profile. We test this by mapping the overall response profile for our animals, finding their maximum and minimum response levels 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.

4.2 Choosing a single assumption

Although making both assumptions together would be ideal, it's likely to be rare in practice. In cases as described above, where the animals were shown to have differing response profiles (with the example of Paddy ranging from 15-350 and Sneezy from 2-180), it must be the case that they differ along at least one of the given dimensions of welfare experience or level of response. And, as we saw, which assumption we make will lead us to a different result – making the first assumption tells us that the otters are enjoying their treat equally, while the second tells us that Paddy is experiencing almost twice as much pleasure from it as Sneezy. It matters a lot in these cases which assumption we choose, so how do we decide between them?

The answer to this is going to depend primarily on context. The justifications provided will support either assumption with varying strength, depending on contextual details such as the indicators used, or the relatedness of the animals. We can look either for which assumption holds the stronger justification or use additional methods to decide between the assumptions. 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 observed differences are a result of different response levels, where scope of welfare experience is held fixed (Assumption 1). 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 and 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 different responses reflect different levels of experience, where response profiles are likely to be similar (Assumption 2). In cases where we see homology in brain structures and processes (as discussed in Section 3), we may be more likely to accept that there will be similarity in experience (Assumption 1). There is important future work to be done here, linking the conditions for evolution of subjective experience and particular indicator responses, to what they can tell us about welfare of individuals and comparisons within and between species. The more we know about the conditions under which subjective experience arose, the mechanisms which create it, and how welfare experience links to changes in the measurable indicators, the better we will become at determining when and how the assumptions will hold.

Another way of deciding between the assumptions is to look for convergence between different types of 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 (Assumption 2). 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 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.

4.3 Making neither assumption

In many cases then, we may be able to justify using one or the other of our similarity assumptions, depending on contextual features of the case, the relationships between the animals, and the nature of the indicators used. However, in a large set of cases even this will be unavailable to us, as neither assumption will hold. These are cases in which the animals are not similar enough in the relevant respects, and they are likely to vary across both dimensions (experience and response profile) at once.

The similarity assumptions rely on justifications of analogy and shared evolutionary history that will only hold for animals which share the relevant similarities of 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, 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 quite 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?

5. Alternative methods

We've seen that most welfare comparisons proceed through 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. For now, I'll take it as likely to be safe only for cases of the same or closely related species. This will typically be sufficient to allow intraspecies and intraindividual comparisons as described in Section One, but not many of the interspecies comparisons, which are some of the most important applications. In this section I'll describe the alternative methods we may use in these cases where similarity assumptions fail, discussing two options – use of sentience proxies, and recourse to other ethical or distributive principles.

5.1 Sentience proxies

One method for trying to make comparisons is the use of alternative proxies for sentience or welfare capacity, typically some form of neural or cognitive complexity (e.g. Budolfson and Spears 2020). There are different possible proxies, including brain size, number of neurons, and connective complexity. It's now common to take measures of complexity rather than sheer size as better potential indicators for sentience (Proctor 2012). This relies on the idea that cognitive complexity might underlie the capacity to experience certain ranges of subjective welfare, such as if we think that cognitive sophistication allows for enjoyment of an expanded range of goods that more simple minds cannot access (Višak 2017). While it makes sense to think that that size or complexity of the brain may relate to the potential breadth and depth of subjective experience, this is an empirical claim – one that could, at least in theory, be tested – and it's not at all clear in advance whether it will turn out to be the case.

The relationship between brain size and cognitive complexity is questionable (Logan et al., 2018) and current distribution of cognitive complexity seems to be poorly correlated with brain size (Paul et al. 2020). This undermines the case for a relationship between brain size and sentience capacity. This is even more true when we consider that there are likely to be multiple dimensions to conscious experience (Birch, Schnell, and Clayton 2020) and we are interested here only in the *affective dimension* that is relevant to welfare experience. Size or complexity of brain regions related to affective processing (such as the periaqueductal grey (Learmonth 2020)) may be more appropriate measures but this is an open question. In practice, the link between cognitive complexity and intensity of affective experience is likely to be highly complex (Yeates 2012). As it stands, we don't currently have strong reason to think that cognitive complexity correlates with welfare capacity (Browning 2019). Additionally, sentience proxies are difficult to validate, without running into the problems already described in this paper (see also Browning and Veit 2020).

Future work in understanding the mechanisms for production of sentient experience may provide answers leading to valid proxies, but this is not yet the case, until our sentience research progresses sufficiently 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, McHaffie, and Yen 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 and 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.

Sentience proxies may sometimes be the best available option in cases where we have no other information to go on – they are at least empirically informed, even if they rely on some untested assumptions. However, they must be applied with caution, and with their current limitations acknowledged. Alternatively, in these situations in which we do not think the similarity assumptions hold, we may instead shift the problem and look at alternative methods of decision-making that do not involve direct comparisons between welfare.

5.2 Giving up on comparisons

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).

Instead, we can switch to a comparison of moral value, rather than directly of sentience or welfare capacity. 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. Perhaps the most promising method then 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 we apply such a principle, we can then use something like the zero-one rule and assign the 0-1 scores to the maximum and minimum welfare levels of each animal, based on the relevant *moral importance* of these states rather than their comparative empirical value. Using these, we can then make our decisions through assessing different actions based on how far up their own scale each species might move – for instance, we might prefer the lungfish furniture if we have a 20% increase for each of the 10 animals, where the lion only has a 30% increase for the single individual. Instead of comparing improvements on a single objective scale of absolute 'welfare units', we would instead compare how much difference they make *relative to the individuals under consideration* and rate them this way. This situation, while not empirically ideal, may capture much of what is important when making such decisions, such as giving equal weight to the interests of different individuals.

There are many potential ways of assigning moral value, which I'll not describe here, but in the absence of the possibility of making determinate comparisons, an equal consideration of individuals view is perhaps the best we can do. 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. 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.

6. 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 specific solution; there's no 'magic bullet' to solve this complex problem. Rather than answering the question of whether we can make comparisons, this can instead be seen as attempting to provide a justification for the practice, and 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 cannot justify such assumptions and instead may need to use different ethical or distributive principles to make the decisions in which we would otherwise want to use comparisons.

In the end, none of the available methods for making welfare comparisons are ideal. They rely on background assumptions for which there's currently insufficient 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 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, Kent C, and Morten L Kringelbach. 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 *The Oxford Handbook of Philosophy of Economics*, edited by Harold Kincaid and Don Ross, 540–59. New York: Oxford University Press.

Birch, Jonathan, Alexandra K. Schnell, and Nicola S. Clayton. 2020. "Dimensions of Animal Consciousness." *Trends in Cognitive Sciences*. 24(10): 789-801

Boccia, Maria L., Mark L. Laudenslager, and Martin L. Reite. 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): 250–57.

Boissy, Alain, Gerhard Manteuffel, Margit Bak Jensen, Randi Oppermann Moe, Berry Spruijt, Linda J. Keeling, Christoph Winckler, et al. 2007. "Assessment of Positive Emotions in Animals to Improve Their Welfare." *Physiology & Behavior* 92(3): 375–97.

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.

Browning, Heather. 2019. "What Should We Do about Sheep? The Role of Intelligence in Welfare Considerations." *Animal Sentience* 4(25): 23.

———. forthcoming. "The Measurability of Animal Welfare." *Journal of Consciousness Studies*.

Browning, Heather, and Walter Veit. 2020. "The Measurement Problem of Consciousness." *Philosophical Topics*. 48(1): 85-108

Budolfson, Mark, and Dean Spears. 2020. "Quantifying Animal Well-Being and Overcoming the Challenge of Interspecies Comparisons." In *The Routledge Handbook of Animal Ethics*, edited by Bob Fischer, 92–101. New York, NY: Routledge.

Coghill, Robert C., John G. McHaffie, and Yi-Fen Yen. 2003. "Neural Correlates of Interindividual Differences in the Subjective Experience of Pain." *Proceedings of the National Academy of Sciences* 100(14): 8538–42.

Davidson, Richard J. 1992. "Emotion and Affective Style: Hemispheric Substrates." *Psychological Science* 3(1): 39–43.

Duncan, I. J. 2002. "Poultry Welfare: Science or Subjectivity?" *British Poultry Science* 43(5): 643–52.

Elster, Jon, and John E. Roemer. 1991. *Interpersonal Comparisons of Well-Being*. Cambridge: Cambridge University Press.

Fleurbaey, Marc, and Peter J. Hammond. 2004. "Interpersonally Comparable Utility." In *Handbook of Utility Theory: Volume 2 Extensions*, edited by Salvador Barberà, Peter J. Hammond, and Christian Seidl, 1179–1285. Boston: Springer US.

Ginsburg, Simona, and Eva Jablonka. 2019. *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness*. Cambridge: MIT Press.

Godfrey-Smith, Peter. 2017. "Animal Evolution and the Origins of Experience." In *How Biology Shapes Philosophy*, edited by David Livingstone Smith, 51–71. Cambridge: Cambridge University Press.

Griffin, James. 1986. *Well-Being: Its Meaning, Measurement and Moral Importance*. Oxford: Clarendon Press.

Harsanyi, John C. 1955. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *Journal of Political Economy* 63(4): 309–21.

Izzo, Genevieve N., Meredith J. Bashaw, and John B. Campbell. 2011. "Enrichment and Individual Differences Affect Welfare Indicators in Squirrel Monkeys (Saimiri Sciureus)." *Journal of Comparative Psychology* 125(3): 347–52.

Learmonth, Mark James. 2020. "The Matter of Non-Avian Reptile Sentience, and Why It 'Matters' to Them: A Conceptual, Ethical and Scientific Review." *Animals* 10(5): 901.

List, Christian. 2003. "Are Interpersonal Comparisons of Utility Indeterminate?" *Erkenntnis* 58(2): 229–60.

Manteca, X., and J. M. Deag. 1994. "Individual Variation in Response to Stressors in Farm Animals: Implications for Experimenters." *Animal Welfare* 3(3): 213–18.

Mason, Georgia. 2010. "Species Differences in Responses to Captivity: Stress, Welfare and the Comparative Method." *Trends in Ecology & Evolution* 25(12): 713–21.

Mason, Georgia, and Michael Mendl. 1993. "Why Is There No Simple Way of Measuring Animal Welfare?" *Animal Welfare* 2 (4): 301–19.

Mellor, D. J. 2016. "Updating Animal Welfare Thinking: Moving beyond the 'Five Freedoms' towards 'A Life Worth Living.'" *Animals* 6 (3): 21.

Nielsen, Christopher S., Donald D. Price, Olav Vassend, Audun Stubhaug, and Jennifer R. Harris. 2005. "Characterizing Individual Differences in Heat-Pain Sensitivity." *Pain* 119(1–3): 65–74.

Panksepp, Jaak. 2011. "Cross-Species Affective Neuroscience Decoding of the Primal Affective Experiences of Humans and Related Animals." *PLOS ONE* 6(9): e21236.

Paul, Elizabeth S., Shlomi Sher, Marco Tamietto, Piotr Winkielman, and Michael T. Mendl. 2020. "Towards a Comparative Science of Emotion: Affect and Consciousness in Humans and Animals." *Neuroscience & Biobehavioral Reviews* 108(January): 749–70.

Pettigrew, Richard. 2019. *Choosing for Changing Selves*. Oxford: Oxford University Press.

Proctor, Helen. 2012. "Animal Sentience: Where Are We and Where Are We Heading?" *Animals* 2(4): 628–39.

Seth, Anil K., Bernard J. Baars, and David B. Edelman. 2005. "Criteria for Consciousness in Humans and Other Mammals." *Consciousness and Cognition* 14(1): 119–39.

Singer, Peter. 2016. "Afterword." In *The Ethics of Killing Animals*, edited by Tatjana Višak and Robert Garner, 229–36. Oxford, New York: Oxford University Press.

Strausfeld, Nicholas J., and Frank Hirth. 2013. “Deep Homology of Arthropod Central Complex and Vertebrate Basal Ganglia.” *Science* 340(6129): 157–61.

Višak, Tatjana. 2017. “Cross-Species Comparisons of Welfare.” In *Ethical and Political Approaches to Nonhuman Animal Issues*, edited by Andrew Woodhall and Gabriel Garmendia da Trindade, 347–63. Cham: Springer International Publishing.

Wimsatt, William C. 2007. *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge: Harvard University Press.

Yeates, J. W. 2012. “Brain-Pain: Do Animals with Higher Cognitive Capacities Feel More Pain? Insights for Species Selection in Scientific Experiments.” In *Large Animals as Biomedical Models: Ethical, Societal, Legal and Biological Aspects*, edited by Kristin Hagen, Angelika Schnieke, and Felix Thiele, 24–46. Bad Neuenahr-Ahrweiler: Europäische Akademie.

Zuolo, Federico. 2017. “Equality, Its Basis and Moral Status: Challenging the Principle of Equal Consideration of Interests.” *International Journal of Philosophical Studies* 25(2): 170–88.