*Utilitas* (2024), **36**, 50–63
doi:10.1017/S0953820823000286

CAMBRIDGE UNIVERSITY PRESS

ARTICLES

# Intersubstrate Welfare Comparisons: Important, Difficult, and Potentially Tractable

Bob Fischer[1] and Jeff Sebo[2]

[1]Department of Philosophy, Texas State University, San Marcos, TX, USA and [2]Department of Environmental Studies, New York University, New York, NY, USA
**Corresponding author:** Jeff Sebo; Email: jeffsebo@nyu.edu

**Abstract**
In the future, when we compare the welfare of a being of one substrate (say, a human) with the welfare of another (say, an artificial intelligence system), we will be making an intersubstrate welfare comparison. In this paper, we argue that intersubstrate welfare comparisons are important, difficult, and potentially tractable. The world might soon contain a vast number of sentient or otherwise significant beings of different substrates, and moral agents will need to be able to compare their welfare levels. However, this work will be difficult, because we lack the same kinds of commonalities across substrates that we have within them. Fortunately, we might be able to make at least some intersubstrate welfare comparisons responsibly in spite of these issues. We make the case for cautious optimism and call for more research.

**Keywords:** AI welfare; animal welfare; welfare comparisons; sentience; moral status

## 1. Introduction

When we judge that one person is better off than another, we make an interpersonal welfare comparison. For instance, when we judge that those who have adequate food are better off than those who are starving, we compare welfare across persons. These comparisons are essential to the assessment of many actions and policies. For instance, if we need to determine whether to prioritize helping one group or another, then their welfare levels matter as one factor among many. And while interpersonal welfare comparisons are difficult, researchers have developed tools (such as surveys that ask for self-reports) that allow us to make these comparisons with, if not full reliability, then at least enough reliability to be useful for some purposes.[1]

When we judge that a member of one species is better off than a member of another, we make an inter*species* welfare comparison. For instance, when we judge that cats who have adequate food are better off than dogs who are starving, we compare welfare across species. Again, such comparisons are essential to the assessment of many actions and policies. If we need to determine whether to prioritize helping the cats or the dogs,

[1]For more on the reliability of self-reported welfare data, see Sandvik et al. (1993) and Caputo (2017).

then their welfare levels matter as one factor among many. And while interspecies comparisons are harder than intraspecies comparisons (in part because we have much more in common within species than across them), researchers are currently developing new tools that make this problem tractable.

When we judge that a being of one substrate (say, a carbon-based animal) is better off than a being of another substrate (say, a silicon-based robot), we make an inter*substrate* welfare comparison. We may not need to make such comparisons now. However, it is plausible that we will need to make them in the future. If or when we do, we will need to determine whether intersubstrate comparisons are tractable. If they are tractable, then they will probably be harder than intrasubstrate comparisons in many respects; once again, researchers will need to develop new tools for making them. If they are not tractable, then researchers will need to develop new tools for making good decisions in the absence of such comparisons.

This paper makes the case that intersubstrate welfare comparisons are important, difficult, and potentially tractable. In a world that contains a vast number and wide range of potentially sentient or otherwise significant beings of different substrates, moral agents will need to be able to include all these beings in our impact assessments and policy decisions in an integrative manner. The bad news is that developing tools for making intersubstrate welfare comparisons will be challenging, as we lack the same kinds of physical and evolutionary "common denominators" across substrates that we have within them. But the good news is that we might be able to develop these tools in spite of these challenges.

Section 2 discusses the importance of intersubstrate welfare comparisons by explaining why intrasubstrate comparisons matter and extrapolating that intersubstrate comparisons will matter for similar reasons. Section 3 discusses the difficulty of intersubstrate welfare comparisons by explaining why our current tools for making welfare comparisons appear to be inapplicable in this context. Section 4 makes a case for the potential tractability of intersubstrate welfare comparisons by presenting four considerations that support a presumption of tractability. We close with a discussion about timing, making the case that this problem is not only important but also urgent; so, we should start developing solutions now.

## 2. Why intersubstrate welfare comparisons are important

Welfare comparisons are essential to ethics and policy decisions. They are especially important when our actions or policies are likely to benefit some individuals while harming others. Such conflicts require a principled resolution. And while many factors may be relevant, one important factor is how much our actions or policies will benefit or harm different stakeholders, in expectation. Welfare comparisons can also be important when our actions or policies are likely to benefit everyone affected (or, at least, are likely to not harm anyone affected). For example, if we have a responsibility to ensure that our actions or policies benefit the least well off among us, then we need to make welfare comparisons to identify those individuals.

We take for granted that welfare comparisons are important within our own species. When we have conflicting interests, we sometimes need to compare the strengths of those interests to resolve those conflicts. Suppose that a doctor is deciding which patient to treat. Bob has a severed artery, is expressing agonizing suffering, and is likely to die without treatment. Jeff has a papercut, is expressing mild discomfort, and is not at risk of dying from this injury. Assuming that all else is equal, the doctor should treat Bob, as Bob has more at stake than Jeff in this case. In particular, Bob is *worse off* than Jeff at

https://doi.org/10.1017/S0953820823000286 Published online by Cambridge University Press

present (in virtue of the relative intensity of his suffering) and has the potential to be *much* worse off in the future (if he dies prematurely).

Interpersonal welfare comparisons are now an essential part of law and policy for similar reasons. Governments need to manage populations of thousands, millions, or even billions of people. And in many cases, they need to make decisions that involve trade-offs within and between these populations. For example, at present the five leading causes of death in the USA are reportedly heart disease, cancer, COVID-19, accidents, and strokes (CDC 2023). How can the U.S. government make a principled decision about which problems to prioritize? At least in part, they can do so by examining the scale of each problem: How many people are impacted by each problem in total and how much are they impacted by each problem on average?

Fortunately, researchers have developed a wide range of tools for making these comparisons at scale in a principled way. For example, if we want to compare how much pain Bob and Jeff are experiencing, then we can ask them to rank their pain on a scale from 1 to 10. If Bob selects 10 and Jeff selects 1, then we have at least *some* evidence that Bob is suffering more than Jeff in this case. Likewise, we can ask Bob and Jeff how much they would be willing to pay to reduce their pain. If, given access to equal resources, Bob reports that he would be willing to pay $100 and Jeff reports that he would be willing to pay only $1, then we once again have at least *some* evidence that Bob is suffering more than Jeff in this case.

Granted, these comparisons are not perfectly reliable. We can easily make mistakes about what others are feeling, including by overestimating or underestimating the strength of their interests. For instance, some people tend to overstate their pain while other people tend to understate their pain (Jamner and Schwartz 1986; Miller and Newton 2006). Additionally, health providers systematically underestimate patients' pain (Seers et al. 2018), and their interpretations of patients' testimony appear to be sensitive to racism (Staton et al. 2007; Trawalter and Hoffman 2015; Trawalter et al. 2012), sexism (Paganini et al. 2023; Robinson and Wise 2003; Zhang et al. 2021), ableism (McGuire et al. 2010), and other such forces.

Still, we can mitigate these risks by correcting for bias and taking other precautionary measures. We can also conduct sensitivity analyses by asking whether these mistakes would change our decisions. And of course, a lot depends on the pros and cons of alternative decision procedures since, in many cases, even unreliable welfare comparisons might be better than none at all. For example, when the doctor considers all the reasons why Bob and Jeff might be offering unreliable self-reports, she might conclude that relying on these self-reports is a risk. But if she needs to make a decision right now and her only options are to, say, rely on self-reports or flip a coin, then relying on self-reports might still be best all things considered.

Increasingly, we recognize that interspecies welfare comparisons are important too – and for many of the same reasons. Consider a variant of the case involving Bob and Jeff. Rob, a dog, has a severed artery, is expressing (or, at least, appears to be expressing) agonizing suffering, and is likely to die without treatment. And Jeff, a human, has a papercut, is expressing (or, at least, appears to be expressing) mild discomfort, and is not at risk of dying. In this case, assuming that all else is equal, it seems plausible that a medical professional should prioritize Rob over Jeff. Granted, we might feel somewhat more uncertain in the Rob–Jeff case than in the Bob–Jeff case, but insofar as we expect that Rob has more at stake, that factor seems relevant to our decision.

As Budolfson et al. (2023) argue, interspecies welfare comparisons should become standard in law and policy in the same kind of way that interpersonal welfare

https://doi.org/10.1017/S0953820823000286 Published online by Cambridge University Press

comparisons are. After all, humans currently kill *billions* of captive animals and *trillions* of wild animals each year for food alone, not including insects.[2] Humans also neglect countless animals during disease outbreaks, fires, floods, and other disasters, even when it would be relatively inexpensive to help them (Green 2019; Sebo 2022). For us to assess the ethics of harming and neglecting animals in these ways, we need to ask a variety of questions, including how the harms that particular practices cause animals compare with the benefits that they provide humans.

Researchers are currently developing tools that we can use to make these comparisons. For example, Sebo (2018) adapts principles of risk to make welfare estimates under uncertainty. Budolfson and Spears (2019) adapt formal tools from economics to make interspecies comparisons. Browning (2023) argues that key similarity assumptions allow for interspecies comparisons in some cases. Veit (2023) argues that life-history differences can track phenomenological differences. Fischer (2024) argues that we can use a variety of empirical proxies to make interspecies comparisons. And Višak (2023) argues that animals have equal capacities for welfare, thereby removing a variable from interspecies comparisons.

Granted, interspecies comparisons are less reliable than interpersonal comparisons. We lack the ability to use verbal self-reports to compare impacts across species on a common scale. And while researchers are developing novel tools for making interspecies comparisons, many questions remain about which tools are best. Simple proxies for welfare capacities like neuron counts and lifespans are clearly unreliable for many purposes (Shriver 2023). Yet complex proxies introduce disagreements and uncertainties that are difficult to resolve (Fischer 2024). And of course, speciesism is at least as influential, if not more influential, than human oppressions that limit our ability to estimate others' welfare in many contexts.

Fortunately, as in the intraspecies case, we can mitigate these risks by taking precautionary measures. We can also use sensitivity analyses to assess our welfare comparisons, and we can note that even unreliable comparisons might still be reliable enough for many purposes. Granted, even with these precautionary measures in place, we might not be able to achieve the same level of precision with interspecies comparisons as with intraspecies ones. But in cases where we need to decide between, say, minor burdens on small human populations and major burdens on large nonhuman populations, we might not *need* to achieve the same level of precision to clarify which population has more at stake in the aggregate.

In the future, intersubstrate welfare comparisons will be important too. Consider another variant of the case involving Fob and Jeff. Fob, a virtual human who exists in the future, is similar to Jeff, a human who likewise exists in the future. The only difference is that Fob is silicon-based and exists in virtual space. Unfortunately, Fob has a (virtual) severed artery, with everything that entails. And Jeff, as usual, has a (physical) papercut, with everything that entails. In this case, we may or may not have the intuition that Fob has more at stake than Jeff. But if we take there to be at least a non-negligible chance that Fob is indeed suffering, that his welfare is negative, then we should at least recognize the question as an important one.

Moving forward, moral agents (humans as well as, eventually, artificial intelligence (AI) systems) are likely to face this kind of question on a regular basis. Humans already

[2]The Food and Agriculture Organization (2023) provides data on the number of farmed land animals killed for food, reporting that over 70 billion are slaughtered globally each year. Meanwhile, an estimated 0.79–2.3 trillion wild fish are caught and slaughtered annually, not including bycatch (Mood et al. 2023).

https://doi.org/10.1017/S0953820823000286 Published online by Cambridge University Press

use AI systems in a variety of ways. We use them as assistants at work, as companions at home, and as allies or adversaries in video games. And in the future, we might build vast digital worlds for research, education, or entertainment. Indeed, given the possibilities available in digital space, the future could contain a vaster number and wider range of non-biological beings than biological beings, in the same kind of way that the present contains a vaster number and wider range of invertebrates than vertebrates. In such a world, intersubstrate welfare comparisons will be essential.

Thus, the task we face is, once again, to develop tools that we can use to make these comparisons. When policymakers face decisions that involve tradeoffs between humans, animals, and AI systems, they need a framework for comparing our interests on a common scale. As with the interpersonal and interspecies cases, we might need to consider many factors – including rights, virtues, and relationships – before we can know what to do. But insofar as expected welfare impacts will be among these factors (and will shape how we assess the other factors), we need a way to estimate the likelihood that non-biological beings can have welfare states and how much welfare they can have relative to biological beings, if any at all.

However, we can expect that intersubstrate welfare comparisons will be harder than intrasubstrate comparisons, in the same kind of way that interspecies comparisons are harder than intraspecies comparisons (and, for that matter, interpersonal comparisons are than intrapersonal comparisons). Specifically, the tools that researchers are developing for making interspecies comparisons might fail to apply to intersubstrate comparisons, in the same kind of way that the some of the tools that researchers developed for making interpersonal comparisons fail to apply to interspecies comparisons. And we might have biases against beings of other substrates in the same kind of way that we do against members of other species.

The question, then, will be exactly how difficult intersubstrate comparisons are and whether we can make these comparisons tractable. Might we be able to use precautionary measures like sensitivity analyses to assess intersubstrate comparisons? And might we face situations where even unreliable intersubstrate comparisons are better than nothing? Suppose that a physical house containing a biological ant and a virtual house containing a virtual human are both burning down, and an unusually positioned firefighter has time to save either being but not both. Suppose further that the firefighter can either flip a coin or use cognitive complexity and longevity to break the tie. Are these proxies reliable enough to be useful in this case?

## 3. Why intersubstrate welfare comparisons are difficult

Some of the difficulties that we anticipate for intersubstrate welfare comparisons are similar to difficulties that we experience with intrasubstrate welfare comparisons. These include difficulties involved with selecting theories of welfare and placing welfare ranges for different kinds of subjects on a common scale. Other difficulties that we anticipate are different, such as the lack of a physical or evolutionary common denominator between beings of different substrates, though we might also see these difficulties as amplifications of ones that we face in the intrasubstrate context. In this section we explain how these difficulties arise in the intrasubstrate case and how they might extend, in amplified form, to the intersubstrate case.

We can start by considering two difficulties that we clearly face in both the intrasubstrate context and the intersubstrate context. The first concerns how to select a theory of welfare. Researchers continue to face disagreement and uncertainty about the basis for

https://doi.org/10.1017/S0953820823000286 Published online by Cambridge University Press

welfare. Some people think that welfare is a matter of experiential states like pleasure and pain. Others think that welfare is a matter of motivational states like desires and preferences. Others think that welfare is a matter of life processes like survival and flourishing. Others think that welfare is a matter of an objective list of goods, where this list might vary both within and across species and substrates depending on the form of life that particular kinds of welfare subjects have.[3]

Each theory of welfare has different implications for welfare comparisons. For example, if welfare is a matter of experiential states, then welfare comparisons might involve comparing how much pleasure, pain, and other such experiential states particular subjects can have. If welfare is a matter of motivational states, then welfare comparisons might involve comparing how much satisfaction, frustration, and other such motivational properties particular subjects can have. If welfare is a matter of life processes, then welfare comparisons might involve comparing how much particular beings can flourish – and might, as Korsgaard (2018) and others argue, be more likely to be incomparable across forms of life. And so on.

Thus, insofar as disagreement and uncertainty remain about which theory of welfare is correct, disagreement and uncertainty will remain about whether and how to make welfare comparisons both within and across species and substrates. For this reason, welfare comparisons will likely require the application of principles of both normative and descriptive uncertainty. For instance, researchers might need to estimate how likely each theory of welfare is to be correct; then estimate how much welfare, if any, particular beings can have according to each theory of welfare; and then aggregate these estimates to produce a general estimate about whether and to what extent particular beings can have welfare. This will, of course, be difficult to do.

However, while welfare comparisons might be difficult in light of this issue, they are not necessarily impossible. We do have tools for addressing both normative and descriptive uncertainty that we can apply in many contexts. When we feel uncertain about whether a particular moral theory is correct, or when we feel uncertain about whether a particular action will be helpful or harmful, we can apply precautionary principles, expected-value principles, or other such principles to decide what to do. Granted, we might not always make the right decisions. But, in many contexts, we can still make better decisions with these tools than without them. And uncertainty about welfare comparisons may well be one of those contexts.

The second difficulty that we clearly face in both the intrasubstrate and the intersubstrate contexts concerns how to place welfare ranges for different kinds of subjects on a common scale. To appreciate this issue, consider a simple model for making welfare comparisons that assigns each subject a welfare range of −1 to 1, which means that each subject's worst welfare state corresponds to −1 and that each subject's best welfare state corresponds to 1. On this model, if we can make welfare *assessments* – that is, if we can assign a number between −1 and 1 to each subject's welfare states – then we can also make welfare *comparisons* – that is, we can compare each subject's numbers to estimate which subject is better or worse off overall.

However, we might not be warranted in accepting such a model, since different subjects might have different welfare ranges. For example, it might be that some subjects

[3]Derek Parfit, in *Reasons and Persons* (1984: 493), was the first to delineate the now-standard classification of welfare theories as a matter of experiential states, desire satisfaction, or objective lists of goods. There are other theories, however, including the life processes account described by Christine Korsgaard (2018).

https://doi.org/10.1017/S0953820823000286 Published online by Cambridge University Press

can have more intense hedonic experiences than others (i.e., that some subjects have hedonic welfare ranges that go higher than 1 or lower than −1). It might also be that welfare involves more than experience, and that other determinants of welfare can vary too (i.e., that even if two subjects have the same *hedonic* welfare ranges, they might not have the same welfare ranges *tout court*). In either case, we might need to reject the idea that each subject's worst welfare state does, in fact, correspond to −1 and that each subject's best welfare state does, in fact, correspond to 1 on the scale.

Of course, if different subjects do have different welfare ranges, this does not necessarily mean that we would be wrong to assume that all welfare states correspond to a point between −1 and 1 (assuming that all welfare states are comparable). After all, which numbers we select for the worst and best possible welfare states are arbitrary. Instead, it means that we would be wrong to assume that each subject's worst possible welfare state corresponds to −1 and that each subject's best possible welfare state corresponds to 1. For example, it might be that the worst and best possible welfare states for an elephant are relatively close to −1 and 1, respectively, but that the worst and best possible welfare states for an ant are relatively close to 0.

But once again, while welfare comparisons might be difficult in light of this issue, they are not necessarily impossible. If we can determine where each subject's worst and best welfare states are between −1 and 1, then we can once again make welfare assessments and comparisons. Suppose that we estimate that the elephant's welfare range is −0.9 to 0.9 and that the ant's welfare range is −0.009 to 0.009. Now, suppose that we estimate that a particular elephant is 10% as badly off as they can possibly be and that a particular ant is 50% as badly off as they can possibly be. In that case, we can estimate that the elephant is worse off than the ant overall, since 10% of −0.9 is −0.09 whereas 50% of −0.009 is −0.0045, and −0.09 is worse than −0.0045.

However, we can now consider a difficulty for intersubstrate welfare comparisons that appears distinctive (though, as we will see, we might also regard it as an extension of difficulties that we face in the intrasubstrate case). This difficulty concerns the apparent lack of relevant common denominators across substrates. To see how this problem arises, suppose for the sake of discussion – as we will for most of the rest of this paper – that welfare is a matter of experiential states like happiness and suffering and that some subjects can experience more intense happiness and suffering than others. In that case, as we have seen, comparing welfare across species requires estimating the intensity of these experiential states across species.

The question is: How can we select numbers for the worst and best welfare states for different kinds of subjects, given that we lack direct access to their experiences? The answer (if, indeed, there is an answer) is that we need to select observable proxies for their unobservable experiences. For example, suppose we have reason to believe that subjects with more informational processing power can have more intense experiential states than subjects with less informational processing power. (We are not defending this claim; we are simply considering one possible proxy, the quality of which we leave open.) In this case, we can use the range of informational processing power as a proxy for the intensity of experiential states.

Whichever proxies we select, what justifies this method (insofar as this method is justified) is the assumption that these proxies reliably track subjects' welfare states. And what justifies this assumption (insofar as this assumption is justified) is the assumption that subjects' welfare states have broadly similar structures, functions, and origins. Granted, there are many differences across species and interspecies welfare comparisons are difficult to make reliably in light of these differences. But there are also

https://doi.org/10.1017/S0953820823000286 Published online by Cambridge University Press

many similarities across species and interspecies welfare comparisons are at least *possible* to make, albeit unreliably, in light of these similarities: the similarities allow us to place welfare ranges on a common scale.

However, this assumption may not hold in the intersubstrate case. If AI systems do, in fact, have experiential states, their silicon-based welfare states might not only have different origins but also have different structures and functions than our carbon-based welfare states. And once our welfare states have different origins, structures, and functions, we might not be warranted in assuming that, say, a given amount of informational processing power corresponds to, say, a given amount of positive or negative conscious experience. So, even if we can use proxies like informational processing power to place welfare states on a common scale in the interspecies case, we might not be able to do the same in the intersubstrate case.

The problem that we face, then, is that intersubstrate welfare comparisons are both important and difficult. As long as AI systems have a non-negligible chance of having the capacity for welfare, we will need to be able to compare expected welfare impacts across substrates in order to know which actions and policies are best (simpliciter or in expectation). And if and when advanced AI systems become moral agents, they will need to be able to do the same. So, we need a method for making welfare comparisons that can include humans, animals, *and* AI systems, not only for altruistic reasons (to ensure that we give proper weight to everyone) but also for self-interested reasons (to ensure that AI systems do the same when the time comes).

But making intersubstrate welfare comparisons will be difficult – or perhaps impossible – given the fundamental differences that exist across substrates. For all the problems we face in the interspecies case, we at least have enough in common for welfare comparisons to be possible on many theories of welfare. But the same might not be true in the intersubstrate case. We may not be able to work out the relationship between, say, a certain kind of pain in humans and the state that reinforces aversion behavior in AI, whatever that state happens to be. In that case, we may need to accept that we cannot directly compare the relative prudential goodness or badness of these welfare states (even if we can recognize them *as* welfare states).

Since it would take a lot of work to develop tools for making intersubstrate welfare comparisons, a first step is to ask to what extent intersubstrate welfare comparisons are promising at all. Insofar as we might eventually be able to make these comparisons with sufficient reliability, we should start developing the tools that might allow us to do so. And insofar as we might *not* be able to make these comparisons with sufficient reliability, we should start developing alternative decision procedures that might allow us to treat humans, animals, and AI systems fairly in spite of our inability to make these comparisons. To what extent should we be making investments in these "optimistic" and "pessimistic" paths at this stage?

## 4. Why intersubstrate welfare comparisons are potentially tractable

Our aim in this section is to defend a modest claim, which is that intersubstrate welfare comparisons are potentially tractable, given the evidence. That is, we should be open to the possibility that these comparisons might or might not be tractable – it would be unreasonable to be either maximally optimistic or maximally pessimistic at this stage – and so we should spend time developing tools for making these comparisons *and* decision procedures for making decisions in the absence of these comparisons as inputs. Since the potential *in*tractability of intersubstrate welfare comparisons is a

https://doi.org/10.1017/S0953820823000286 Published online by Cambridge University Press

given at this stage, this section focuses on four considerations that support the potential tractability of these comparisons.

The first consideration concerns an argument from induction. As the previous sections suggest, we have a long history of going through the following process: we think that particular kinds of welfare comparisons are intractable because our current methods of making welfare comparisons fail to apply to them. We then discover that these welfare comparisons are tractable after all (at least in the sense of being good enough to be worth using for some purposes), by developing new tools that we can use to make them. If, then, we currently think that intersubstrate welfare comparisons are intractable, we should expect to be in a similar situation, and we should expect that a similar discovery is forthcoming.

Consider that some experts have argued that welfare comparisons are intractable even within our own species. After all, every human is different and the problem of other minds applies at this level too. But despite this issue, we have found commonalities within our species that allow us to make at least *some* welfare comparisons with at least *some* confidence; specifically, we assume that all humans have similar welfare ranges in virtue of our shared history, anatomy, and behavior, and so we compare the strength and valence of our interests by examining our history, anatomy, and behavior. While the resulting welfare comparisons might leave a lot to be desired, they are still good enough for many purposes in ethics and policy.

Other experts have argued that welfare comparisons might not be tractable across species, since the commonalities that we use for intraspecies comparisons might not apply in this context. However, experts are now finding commonalities that do apply across species and are creating new methods for making welfare comparisons accordingly. Yes, members of different species might not share *specific* histories, anatomies, or behaviors, but they still share *general* histories, anatomies, and behaviors. And we can use these general commonalities to estimate welfare ranges for particular species and points on these ranges for particular individuals. Again, this might leave a lot to be desired, but it can still be good enough for many purposes.

We might now feel tempted to argue that welfare comparisons might not be tractable across *substrates*, since commonalities that we use for intrasubstrate comparisons might not apply in this context. But we might once again be able to find commonalities that do apply across substrates and create new methods for making welfare comparisons accordingly. For example, there might be general material, structural, or functional similarities between carbon-based and silicon-based systems that allow us to at least *roughly* estimate welfare ranges for AI "species" and points on these ranges for AI systems. Once again, this might leave a lot to be desired, but it might still be good enough for at least some purposes.

The second consideration concerns different kinds of welfare comparison. Welfare comparisons can be made with greater or lesser precision. They can also involve the intensity of welfare states, the valence of welfare states, or both. And of course, whether particular welfare comparisons are sufficiently reliable depends on their intended use. For example, in cases where welfare comparisons need to be both precise and complete (i.e., involve both intensity and valence) to be useful, the bar for tractability is higher and pessimism might be warranted. But in cases where welfare comparisons can be imprecise or incomplete (say, involving valence but not intensity) and still be useful, the bar for tractability is lower and optimism might be warranted.

This point matters because we can expect that imprecise or incomplete welfare comparisons can, in fact, be useful in some cases. For instance, if a house is burning down

and you can save an elephant or an ant but not both, then a precise comparison regarding the intensity of their experiences might not be necessary. Instead, an imprecise comparison (the elephant is *somewhat* likely to suffer *somewhat* more) might be sufficient. And in cases where our goal is to identify Pareto-optimal policies (such that any deviation from these policies that benefits some would harm others), a comparison regarding the intensity of our experiences might not be necessary. Instead, a comparison regarding the valence of our experiences might be sufficient.

The third consideration concerns AI capabilities. By the time we need to make intersubstrate welfare comparisons, AI systems might be able to tell us about their experiences. One of the problems that we face when making interspecies welfare comparisons is that we have no way to validate our welfare measures by self-report. We can develop theories about, say, whether and to what extent ants can experience happiness, suffering, and other such states, but we can never ask ants for confirmation that our theories are correct. Yet this problem might not extend to AI systems, in which case intersubstrate comparisons might be easier than interspecies comparisons in at least one respect, even if they remain harder in other respects.[4]

Of course, one might object that we have no reason to trust AI testimony, either at present or in the future. But while we think that skepticism about AI testimony is reasonable at present, we also think that humans might have at least *some* reason to give at least *some* weight to at least *some* AI testimony in the future. After all, research on AI safety, alignment, and interpretability is ongoing, and if this research goes well, then it might lead to innovations that allow for greater trust between humans and AI systems. If so, then when relatively trustworthy AI systems tell us about the nature and content of their experiential states, we might have at least *some* reason to give at least *some* weight to this testimony, even if only a small amount.

The final consideration concerns the nature of welfare. As we have seen, intersubstrate welfare comparisons might be more difficult according to some theories of welfare than according to others. Our discussion in this paper has focused on the view that welfare is a matter of experiential states, and this view makes intersubstrate (and, indeed, intrasubstrate) welfare comparisons harder, since it implies that the determinants of welfare are not directly observable. But other views – such as the view that welfare is a matter of motivational states or a matter of an objective list of species-specific goods – might make these comparisons easier, since these views might imply that some determinants of welfare *are* directly observable.

Of course, one might object that this point is irrelevant, since welfare *is*, in fact, a matter of experiential states. But while we think that this view is likely correct, we also think that other views have at least a non-negligible chance of being correct. Given the difficulty of moral philosophy and the slow pace of progress in this field, it would be a mistake for any of us to be *certain* that our favorite theory is correct at this stage. Instead, we should give at least some weight to each theory that has at least a non-negligible chance of being correct. And plausibly, we should include at least some theories that make these comparisons easier in that category. Insofar as we do, we should treat these comparisons as at least somewhat tractable.

These considerations support the idea that *some* intersubstrate comparisons are *potentially* tractable in expectation. We conclude that, given the importance of these

[4]For that matter, AI systems might one day be able to "decode" nonhuman animals' sounds and behavior (Bakker 2022; Rutz et al. 2023), allowing us better access to nonhuman animals' experiences and improving our interspecies welfare comparisons.

https://doi.org/10.1017/S0953820823000286 Published online by Cambridge University Press

comparisons, we should attempt to develop tools that might allow us (and other moral agents) to make them in the future. Granted, we might never be able to make intersubstrate comparisons as well as we can make intrasubstrate ones, in the same kind of way that we might never be able to make interspecies comparisons as well as we can make intraspecies ones. But as we have seen, there is at least one respect in which intersubstrate comparisons might be easier. And in any case, even imperfect welfare comparisons can be better than nothing.

To be clear, there are many considerations that support intractability as well. For all we know now, welfare is a matter of experiential states, experiential states within substrates are type identical, and experiential states across substrates are *not* type identical. In that case, the project of comparing the intensity of experiential states within substrates might be tractable, because we would be attempting to compare experiential states of the same type, and these states would, at least in principle, be comparable. But the project of comparing experiential states across substrates might not be tractable, because we would be attempting to compare experiential states of different types, and these states might, even in principle, be incomparable.

Of course, this kind of concern can threaten interspecies welfare comparisons too. For example, Korsgaard (2018) and others argue that each species has a different form of life and each life can be assessed only by the standards set by its form of life. Thus, for instance, we might be able to say that one elephant has a better or worse life than another elephant since these animals have the same form of life; so, we can assess how well or badly their lives are going by reference to the same standard. However, we might not be able to say that an elephant has a better or worse life than an ant, since these animals have different forms of life and we can assess how well or badly their lives are going only by reference to different standards.

But while this kind of concern is reasonable, it does not support abandoning the project of making intersubstrate welfare comparisons at this stage. In general, our effort to succeed at a project – and our tolerance for the uncertainty of success – should increase with the importance of the project. And in this case, the project of developing tools that can allow those in power to make welfare comparisons both within and across species and substrates is *extremely* important. An unfathomable number of biological and non-biological lives could depend on it. So even if we think that it is *much* more likely than not that this project is intractable for these reasons, we should still undertake the project at this stage and see if we can prove ourselves wrong.

To be clear, the key premise in our response to this objection is *not* that insofar as the project of making intersubstrate welfare comparisons is important, this project is likely to be tractable. That would be a bad inference. Instead, the key premise is that insofar as this project is important, the project can still be worthwhile in expectation even when the probability of success is low. And in our view, the considerations that we presented in favor of potential tractability are more than enough to meet this standard. So, we should allocate research time *both* toward developing tools that might allow us to make these comparisons *and* toward developing decision procedures that might allow us to make good decisions in the absence of such comparisons.

## 5. Conclusion

We have argued that humans should attempt to develop tools for making intersubstrate welfare comparisons so that they can be ready by the time we need them. Of course, one might accept this conclusion but reject the idea that we should start this project anytime

https://doi.org/10.1017/S0953820823000286 Published online by Cambridge University Press

soon. After all, there are likely not any silicon-based welfare subjects yet, whereas there are trillions (if not quadrillions or quintillions) of carbon-based welfare subjects who need our attention. For this reason, one might think that we should focus on developing tools for making intrasubstrate welfare comparisons now and we can then develop tools for making intersubstrate welfare comparisons later on, if and when AI systems are more likely to be welfare subjects.

While we sympathize with this view, the project of developing tools for making intersubstrate welfare comparisons is urgent. First, moral agents should include a being in our moral circle not when this being is *likely* to be a welfare subject, but rather when this being has *a non-negligible chance* of being a welfare subject (Sebo and Long forthcoming). And given how rapidly AI is developing and how much uncertainty we have about relevant facts and values, we might find that some non-biological systems have a non-negligible chance of being welfare subjects soon. We might also find that the number and variety of non-biological systems exceeds the number and variety of biological systems soon after that.

Second, the pace of academic research is generally slow and the project of developing tools for making intersubstrate welfare comparisons will be difficult. We should thus start this project *before* we anticipate needing these tools, not *when* we anticipate needing them. Otherwise we risk reaching the day when AI systems have a non-negligible chance of being welfare subjects and then needing an extra decade or more to develop the basic tools needed for treating them fairly. And during that period, humans might treat many AI systems badly and path dependence might make it harder to change these practices. We should learn from the mistakes that we made with other animals and avoid placing ourselves in this situation with AI systems.

Third, the project of developing tools for making intersubstrate welfare comparisons is not in competition with other urgent projects, such as the project of developing tools for making interspecies welfare comparisons. We can work on multiple projects at once. And, when we work on related projects in an integrative manner, we can develop a big-picture understanding of a general research area that improves our work on each project. In this case, for example, by working on interspecies and intersubstrate welfare comparisons in an integrative manner, we can work toward a maximally general, foundational understanding of how to assess and compare welfare under uncertainty. This understanding will improve our work on each project.

We thus call for philosophers, cognitive scientists, computer scientists, and other scholars to start working on the problem of intersubstrate welfare comparisons *now*. This problem is both important and urgent. It will take time to investigate its tractability, to develop tools that will allow moral agents to make intersubstrate welfare comparisons insofar as the problem *is* tractable, and to develop tools that will allow moral agents to make good decisions in the absence of these comparisons insofar as the problem is *not* tractable. By starting this work now, ideally in collaboration with researchers who work on intrasubstrate welfare comparisons, we can make progress on both issues in an integrative manner, while we still have time.

**Acknowledgments.** The authors thank Toni Adleberg for extensive research and editorial assistance on multiple drafts of this paper, Ben Eggleston and two anonymous referees at *Utilitas* for helpful feedback on the penultimate draft, and the organizers and participants of the 12th Oxford Workshop on Global Priorities Research in June 2023 for helpful discussion about this topic.

**Competing interests.** None.

## References

**Bakker, Karen**. 2022. *The Sounds of Life: How Digital Technology is Bringing Us Closer to the Worlds of Animals and Plants* (Princeton: Princeton University Press).

**Browning, Heather.** 2023. Welfare Comparisons within and across Species. *Philosophical Studies*, **180**.2, 529–51 <https://doi.org/10.1007/s11098-022-01907-1>.

**Budolfson, Mark, Bob Fischer, and Noah Scovronick**. 2023. Animal Welfare: Methods to Improve Policy and Practice. *Science*, **381**.6653, 32–34.

**Budolfson, Mark, and Dean Spears**. 2019. 'Quantifying Animal Well-Being and Overcoming the Challenge of Interspecies Comparisons', in *The Routledge Handbook of Animal Ethics*, ed. Bob Fischer (New York: Routledge), pp. 92–101.

**Caputo, Andrea**. 2017. Social Desirability Bias in Self-Reported Well-Being Measures: Evidence from an Online Survey. *Universitas Psychologica*, **16**.2, 1657–9267.

**CDC**. 2023. Leading Causes of Death, CDC, <https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm> [accessed August 2023].

**Fischer, Bob**, ed. 2024. *Weighing Animal Welfare: Comparing Well-Being Across Species* (New York: Oxford University Press).

**Food and Agriculture Organization of the United Nations**. 2023 Food and Agriculture Data, FAOSTAT, <https://www.fao.org/faostat/en/#home> [accessed August 2023].

**Green, Dick**. 2019. *Animals in Disasters* (Oxford: Elsevier).

**Jamner, L. D., and G. E. Schwartz**. 1986. Self-Deception Predicts Self-Report and Endurance of Pain. *Psychosomatic Medicine*, **48**.3, 211.

**Korsgaard, Christine**. 2018. *Fellow Creatures: Our Obligations to the Other Animals* (New York: Oxford University Press), 3–15.

**Mcguire, Brian, P. Daly, and F. Smyth**. 2010. Chronic Pain in People with an Intellectual Disability: Under-Recognised and Under-Treated. *Journal of Intellectual Disability Research*, **54**, 240–45 <https://doi.org/10.1111/j.1365-2788.2010.01254.x>.

**Miller, Carly, and Sarah E. Newton**. 2006. Pain Perception and Expression: The Influence of Gender, Personal Self-Efficacy, and Lifespan Socialization. *Pain Management Nursing*, **7**.4, 148–52 <https://doi.org/10.1016/j.pmn.2006.09.004>.

**Mood, Alison, Elena Lara, Natasha K. Boyland, and Phil Brooke**. 2023. Estimating Global Numbers of Farmed Fishes Killed for Food Annually from 1990 to 2019. *Animal Welfare*, **32**, e12 <https://doi.org/10.1017/awf.2023.4>.

**Paganini, Gina A., Kevin M. Summers, Leanne ten Brinke, and E. Paige Lloyd**. 2023. Women Exaggerate, Men Downplay: Gendered Endorsement of Emotional Dramatization Stereotypes Contributes to Gender Bias in Pain Expectations. *Journal of Experimental Social Psychology*, **109**, 104520 <https://doi.org/10.1016/j.jesp.2023.104520>.

**Parfit, Derek**. 1984. *Reasons and Persons* (Oxford: Oxford University Press).

**Robinson, Michael E., and Emily A. Wise**. 2003. Gender Bias in the Observation of Experimental Pain. *Pain*, **104**.1, 259–64 <https://doi.org/10.1016/S0304-3959(03)00014-9>.

**Rutz, Christian, Michael Bronstein, Aza Raskin, Sonja C. Vernes, Katherine Zacarian, and Damián E. Blasi**. 2023. Using Machine Learning to Decode Animal Communication. *Science*, **381**.6654, 152–55 <https://doi.org/10.1126/science.adg7314>.

**Sandvik, Ed, Ed Diener, and Larry Seidlitz**. 1993. Subjective Well-Being: The Convergence and Stability of Self-Report and Non-Self-Report Measures. *Journal of Personality*, **61**.3, 317–42 <https://doi.org/10.1111/j.1467-6494.1993.tb00283.x>.

**Sebo, Jeff**. 2018. The Moral Problem of Other Minds. *The Harvard Review of Philosophy*, **25**, 51–70 <https://doi.org/10.5840/harvardreview20185913>.

**Sebo, Jeff**. 2022. *Saving Animals, Saving Ourselves: Why Animals Matter for Pandemics, Climate Change, and Other Catastrophes* (New York: Oxford University Press).

**Sebo, Jeff, and Robert Long**. Forthcoming. Moral Consideration for AI Systems by 2030. *AI and Ethics*.

**Seers, Tim, Sheena Derry, Kate Seers, and R. Andrew Moore**. 2018. Professionals Underestimate Patients' Pain: A Comprehensive Review. *Pain*, **159**.5, 811–18 <https://doi.org/10.1097/j.pain.0000000000001165>.

**Shriver, Adam**. 2023. Why Neuron Counts Shouldn't Be Used as Proxies for Moral Weight. Effective Altruism Forum. <https://forum.effectivealtruism.org/posts/Mfq7KxQRvkeLnJvoB/why-neuron-counts-shouldn-t-be-used-as-proxies-for-moral> [accessed August 2023].

https://doi.org/10.1017/S0953820823000286 Published online by Cambridge University Press

**Staton, Lisa J., Mukta Panda, Ian Chen, Inginia Genao, James Kurz, Mark Pasanen, and others**. 2007. When Race Matters: Disagreement in Pain Perception between Patients and Their Physicians in Primary Care. *Journal of the National Medical Association*, **99**.5, 532–38.

**Trawalter, Sophie, and Kelly M. Hoffman**. 2015. Got Pain? Racial Bias in Perceptions of Pain. *Social and Personality Psychology Compass*, **9**.3, 146–57 <https://doi.org/10.1111/spc3.12161>.

**Trawalter, Sophie, Kelly M. Hoffman, and Adam Waytz**. 2012. Racial Bias in Perceptions of Others' Pain. *PLoS ONE*, **7**.11, e48546 <https://doi.org/10.1371/journal.pone.0048546>.

**Veit, Walter**. 2023. *A Philosophy for the Science of Animal Consciousness* (New York: Routledge).

**Višak, Tatjana**. 2023. *Capacity for Welfare across Species*. Oxford: Oxford University Press.

**Zhang, Lanlan, Elizabeth A. Reynolds Losin, Yoni K. Ashar, Leonie Koban, and Tor D. Wager**. 2021. Gender Biases in Estimation of Others' Pain. *The Journal of Pain*, **22**.9, 1048–59 <https://doi.org/10.1016/j.jpain.2021.03.001>.

**Bob Fischer** is a senior research manager at Rethink Priorities; and an associate professor of philosophy at Texas State University.

**Jeff Sebo** is an associate professor of environmental studies; affiliated professor of bioethics, medical ethics, philosophy, and law; and director of the Mind, Ethics, and Policy Program at New York University.

**Cite this article:** Fischer B, Sebo J (2024). Intersubstrate Welfare Comparisons: Important, Difficult, and Potentially Tractable. *Utilitas* **36**, 50–63. https://doi.org/10.1017/S0953820823000286