Wlodek Rabinowicz¹

Getting Personal—The Intuition of Neutrality Re-interpreted

According to the Intuition of Neutrality, there is a range of wellbeing levels such that adding people with lives at these levels doesn't make the world either better or worse. As lives in the neutral range can be good for those who live them, this intuition is in conflict with one of the main tenets of welfarism; it creates a disparity between what is good for a person and what is impersonally good. Adding a person with a good life needn't make the world better. In "Broome and the Intuition of Neutrality" (2009) I suggested, but did not elaborate, a re-interpretation of the neutral range that would remove the problematic disparity. On this re-interpretation, a life at a level within the neutral range is not merely impersonally neutral; it is also neutral in its personal value: neither better nor worse for its owner than non-existence. Nevertheless, among such personally neutral lives, some might still be personally better or worse than others, provided that they are incommensurable in their personal value with non-existence. In this paper, I explore some of the implications of this 'personalization' of the Intuition of neutrality. In particular, I discuss its worrisome implications for neutral-range utilitarianism (NRU). While NRU was originally proposed as a way to avoid the Repugnant Conclusion, it turns out this conclusion is re-instated on the new interpretation and, contrary to what was suggested in my 2009-paper, it remains repugnant. A related point is that it no longer holds that all personally good lives must be better for a person than personally neutral lives. Nor that all personally bad lives must

¹Department of Philosophy, Lund University, wlodek.rabinowicz@fil.lu.se.

be worse than personally neutral lives. While this might seem strange, it should be accepted. As for the worrisome implications of NRU, these implications do not undermine the personalized Neutrality Intuition itself. The latter might well be retained even if NRU is given up.

*

According to the Intuition of Neutrality, in its axiological version, there is a range of wellbeing levels such that adding people with lives at these levels doesn't make the world either better or worse. On the standard interpretation of this 'neutral range', it extends from the zero level of wellbeing upwards, with the upper limit being some positive, though not very high, wellbeing level. (On the radical interpretation, the range of neutrality still starts at zero but has no upper limit.) Since a life at a positive wellbeing level is thought to be good for the person who lives this life, the Intuition of Neutrality drives a wedge between what is good for a person and what is impersonally good: Adding a person with a life that is good for her but has a wellbeing level within the neutral range doesn't make the world better, even if no one else is negatively affected by the addition. To this extent then, the Intuition comes into conflict with one of the basic tenets of welfarism.

In Rabinowicz (2009) I discussed the Intuition of Neutrality and defended it against John Broome's challenging objections (Broome 2004). I also sketched a particular axiological theory – neutral-range utilitarianism – that incorporates this intuition. While Broome's criticisms did not target the disparity the Intuition brings in between personal and impersonal goodness,² I still thought that it was there the main problem with the Intuition was to be found. Therefore, I also suggested, but did not elaborate in that paper, a re-interpretation of the neutral range that would remove the problematic disparity. On this re-interpretation, a life at a level within the neutral range is not merely impersonally neutral – it does not merely fail to make the world better or worse – but it also is neutral in its personal value: It is neither good nor bad for a person to have such a life. Its wellbeing level is neither positive nor negative. A life at this level is thus neither worth living nor worth not living. To

² Indeed, Broome's own favourite theory in that book, critical-level utilitarianism, also implies such a disparity. In fact, it goes even further in this respect than the Intuition of Neutrality: it implies that what is personally good might be impersonally bad. Adding a person with a life that is good for her makes the world worse if the wellbeing level of the added life, while positive, is lower than the critical level adopted by the theory. To be sure, Broome also suggests that the location of the critical level might well be indeterminate. This would make it indeterminate, for some positive wellbeing levels, whether adding people with lives at those levels makes the world better or worse. Indeterminacy makes the disparity between what's good for a person and what's good for the world less blunt. But it doesn't remove it.

put it differently, it is neither better nor worse for its owner than non-existence. Nevertheless, among such personally neutral lives, some might still be personally better or worse than others. Thus, there might be a range of personally neutral well-being levels – some higher and some lower. This is possible provided that lives at these levels can be incommensurable in their personal value with non-existence. Unlike equal goodness, incommensurability is not a transitive relation, which means that a personally neutral life – a life that is incommensurable with non-existence – can still be personally better than another personally neutral life.

In this paper, I want to explore some of the implications of this ‘personalization’ of the Intuition of neutrality. In particular, while such a move might seem to make neutral-range utilitarianism more plausible, I will argue that the appearances are misleading. For one thing, the resulting theory gets considerably more complicated: We not only need to allow for lives that are incommensurable with non-existence; we also need to allow for lives that are incommensurable with each other. This means that some lives’ wellbeing levels might not be ordinally comparable. Allowing for lives that are incommensurable in personal value makes technical trouble for a utilitarian axiology. It is not even obvious that such an axiology can accommodate incommensurable lives. While this worry, I think, can be dealt with, there is also a substantive issue that neutral-range utilitarianism needs to confront. In my 2009-paper, I pointed out that personalizing the Intuition of Neutrality re-instates the Repugnant Conclusion that neutral-range utilitarianism was supposed to avoid. But I also suggested that this re-interpretation at the same time removes, or at least assuages, the repugnancy of the Repugnant Conclusion. I now think this diagnosis was premature: If the Intuition of Neutrality is re-interpreted on the lines I have suggested, then even a genuinely repugnant conclusion can be re-instated. This poses a challenge to neutral-range utilitarianism. Indeed, the whole landscape of personal value becomes more complicated on this new picture. As suggested above, we need to give up the standard assumption that the wellbeing levels of different lives are linearly ordered. As I am going to show, one of the implications of this change is that it no longer holds that all personally good lives must be better than personally neutral lives. Nor that all personally bad lives must be worse than personally neutral lives. But while such implications of this new approach are surprising, they should be accepted. As for the issues that on this new interpretation arise for the neutral-range utilitarianism, they do not undermine the personalized Neutrality Intuition itself. The latter can be kept even if the neutral-range utilitarianism were given up.

In section 1, I will present the standard interpretation of the Intuition of Neutrality and then move on, in Section 2, to neutral-range utilitarianism (NRU). In section 3, I will describe the personalized version of the Intuition and the way this

personalization affects NRU. The picture I will draw will be essentially the same, although more elaborate, as the one I have sketched in Rabinowicz (2009). Then, in section 4, I will consider how this picture changes and gets more complicated, in ways I didn't envisage in my earlier paper, if one allows that different lives might not be commensurable in their personal value.

One important part of morality is concerned with what is better or worse for people. According to a popular version of this *person-affecting* idea of morality, what is better (worse) must be *better (worse) for* someone.³ However, it is unclear how we are to put this idea to use in non-identity cases, i.e., cases where, depending on what we decide to do, different people will come to exist in the future. Indeed, there seems to be a clear tension between the person-affecting idea and some of our considered judgements about non-identity cases. In at least some non-identity cases we want to say that one outcome is better (or worse) than another in virtue of the wellbeing of people who do not exist in both. For example, we want to say that creating a very unhappy person makes the world worse, other things being equal. But how can we say this, if an outcome is worse only if it worse for someone? In order to comply with a person-affecting morality in this case, we need to show that coming into existence can be worse for a person. But can it really be worse for a person to exist than not to exist, and thus better for her not to exist than to exist? That seems to require that the person would have been better off not existing, which sounds paradoxical.

In this paper, I am going to discuss some recent attempts to ease this tension. According to these attempts, we can stick to a person-affecting morality and still avoid the counterintuitive judgement that no outcome is better or worse in virtue of the wellbeing of people whose existence is contingent on our choice. I shall show that none of these attempts is convincing. That leaves us with only one option: to reject the person-affecting constraint in its current form.

In section 2, I shall say more about non-identity cases, and list the most morally salient ones. In section 3, I shall make more precise what a person-affecting morality amounts to. In section 4, I shall present an argument that spells out the tension between person-affecting morality and our judgements about non-identity cases. The argument's conclusion is that no outcome can be better or worse than another in terms of the well-being of people who do not exist in both. In sections 5 to 10, I shall discuss possible ways to resist this argument while sticking to a person-affecting morality. I shall especially focus on the approach recently defended by the so-called 'Scandinavian existentialists'.⁴ I shall argue that the main problem with

³ See Temkin (1993a), (1993b), and Holtug (1996). The label "Person-Affecting Restriction" was introduced by Glover (1977), p. 66, but see also Narveson (1967).

⁴ See, for instance, Arrhenius & Rabinowicz (2010), (2015), Johansson (2010), and Holtug (2001). See also, Adler (2009), and Adler (2011) for similar ideas. Some seeds for this approach seem to have been planted already in Parfit (1995), appendix G, p. 490.

their approach is that they fail to fully acknowledge what it means to say that an abstract state of affairs has value.

1. The Intuition of Neutrality—standard interpretation

As a normative principle of action, the Intuition of Neutrality goes back to Jan Narveson's famous pronouncement: "We are in favour of making people happy, we are neutral about making happy people." (Narveson 1973) A more guarded, less committal formulation of this ethical Intuition is provided by John Broome in *Weighing Lives*:

We think intuitively that adding a person to the world is very often ethically neutral. We do not think that just a single level of wellbeing is neutral [...].
(Broome 2004, p. 143)

Here, following Broome (ibid. p. 145f), I am going to focus on the axiological version of the Intuition. It can be put as follows:

Intuition of Neutrality: There is a range of wellbeing levels, call it *the neutral range*, such that adding a person with a life at one of these levels, without affecting the wellbeing of anyone else, does not make the world either better or worse.

Broome himself was initially attracted to this Intuition, but then eventually felt compelled to give it up, for several reasons that I am not going to discuss in this paper. But see Rabinowicz (2009) for an extended critical discussion of Broome's objections. For Broome's reply, see Broome (2009).

What is supposed to be the scope of the neutral range? On the moderate interpretation, this range begins at the zero level of wellbeing and extends upwards, to some positive, though not too high, wellbeing level. On the radical interpretation, it extends all the way up, to infinity.⁵ Here, I will focus on the moderate interpretation, according to which the neutral range has an upper bound as well as a lower bound. On this view, adding bad lives – lives at negative wellbeing levels – makes the world worse, while adding excellent lives makes the world better. In what follows, I am

⁵ Cf. Broome (2004, p. 144): "Some people think this range is infinitely wide. They think that a person's existence is neutral, however good her life would be if she did exist. It is not neutral if her life would be bad, so there is a lower boundary to the neutral range. But there is no upper boundary. That is one view. A more moderate view is that the range has both an upper and a lower boundary, but there is nevertheless a range of neutral lives in between."

going to refer to this interpretation of the neutral range as the ‘standard’ one. Though I don’t want to imply by this label that any dominating view has already developed in this area.

By a bad life I mean a *personally* bad life, i.e. a life that is bad for the person who lives it. For that person such a life is ‘worth not living’: It is worse for her than non-existence. I assume that lives at negative levels of wellbeing are bad in this sense.

Analogously, a good life is a life that is personally good, or worth living: it is a life that is better for the person who lives it than non-existence. A life’s wellbeing level is positive iff the life in question is good in this sense.

In line with this interpretation of positive and negative wellbeing levels, the zero level of wellbeing characterizes a life that for the person who lives it has the same value as non-existence.

I will assume that a life’s personal value does not depend on the numerical identity of the person who lives this life. Thus, for anyone who would live this life its personal value would for her be the same. I think it is a reasonable assumption, provided we take a person’s life to contain everything that characterizes this person – not merely what she does and what happens to her, and not just her external circumstances as they change over time, but also her internal, psychological make-up and internal history. Given this all-encompassing conception of a life, it is plausible to assume that, for anyone who would live it, it would have the same value.

Several philosophers have argued that personal value comparisons between one’s life and one’s non-existence make no sense. It makes no sense on their view to suggest that it can be better (or worse) for me to live my life than never to live at all. I disagree; I think that comparisons of this kind do make sense, even though they might be difficult to make. (See Arrhenius and Rabinowicz 2010, 2015; cf. also Johansson 2010 and Holtug 2001 for related views. For challenging objections, see especially Bykvist 2007, 2014.). Here I cannot argue this point, but I can at least say something to allay the immediate worry such a claim might invite.

Consider:

- (i) It is better for John to have the life he has than not to exist.

This statement implies, according to the critic, that

- (ii) If John didn’t exist, it would have been worse for him than to have the life he has.

But, if John didn't exist, nothing *could* have been worse, or better, for him, as there would be no him for whom anything could be better or worse. Non-existents cannot have any properties or stand in any relations. Thus, (ii) is absurd, which shows that such claims as (i) have absurd implications.⁶

In my view, this objection is not justified: Contrary to appearances, (ii) does not follow from (i). (i) and the consequent of (ii) state that a certain relation obtains / would have obtained between three relata: John and two states of affairs, John having the life he has and John's non-existence. Now, a relation can only obtain if all its relata exist. Consequently, it couldn't have obtained if John did not exist. Which means that (ii) must be false: If John did not exist, no relation in which he is one of the relata could have obtained. On the other hand, (i) may well be true. If John does exist, the requisite relation between him and the two states of affairs may well obtain.

But, one might wonder, if John does exist, then what about the state of his non-existence? Does this state exist in such a case? The answer is yes, if we think of states of affairs as abstract objects. As such, they exist even when they do not obtain.⁷ We may therefore conclude that there is no implication from (i) to (ii): (ii) is necessarily false for the reasons that do not apply to (i).

I will therefore continue to account for the personal value of a life in terms of comparisons with non-existence: For any person, a life L is good (bad, neutral) for that person iff it would be better (worse, neither better nor worse) for her to have life L than not to exist at all.

Let us go back to the main thread. The Intuition of Neutrality together with the standard conception of the neutral range imply that adding a good life need not always make the world better. It will not make it better, or worse for that matter, if the wellbeing level of that good life lies within the neutral range. Thus, on this view, there is a striking *disparity* between what is good for a person and what is impersonally good: A life might be good for the person who lives it without being impersonally contributively good – without making the world better. This disparity might well be precisely what attracts some philosophers to the Intuition of Neutrality.

⁶ Cf. Broome (1999, ch. 10, p. 168): "[I]t cannot ever be true that it is better for a person that she lives than that he should never have lived at all. If it were better for a person that she lives than that she should never have lived at all, then if she had never lived at all, that would have been worse for her than if she had lived. But if she had never lived at all, there would have been no her for it to be worse for, so it could not have been worse for her." See also Parfit 1991 [1984], p. 395. In *Weighing Lives*, though, Broome is less categorical: He recognizes that comparing a person's life with her non-existence, in terms of its value for the person in question, might possibly be made sense of after all (see Broome 2004, p. 63).

⁷ Note that only if states of affairs can exist without obtaining can there be any relations between states that do not co-obtain. Incompatible states cannot co-obtain. Thus, not even the relation of incompatibility would obtain between them if they couldn't exist unless they obtained. For if one of them obtains, the other does not.

ty, but at the same time it is a worrying implication from a strictly welfarist point of view. I will return to this issue in Section 3.

How is the neutrality of life additions to be interpreted? If adding a life with a wellbeing level in the neutral range (while keeping the wellbeing levels of everyone else constant) doesn't make the world either better or worse, will the world with such an added life be equally as good as the original world? Or will it instead be *incommensurable* with the original world – neither better or worse nor equally as good? It is easy to prove that incommensurability is the only viable alternative.

As a preparation for the proof let us note that the neutral range is supposed to contain more than just one wellbeing level. Furthermore, according to the standard conception of the neutral range, the wellbeing levels are linearly ordered, from higher to lower. Consequently, it follows that *for very wellbeing level m in the neutral range there is at least one level in that range that is higher or lower than m* . (While the linear ordering assumption will be criticized in section 4, the italicized claim won't be questioned. And it is only this weak claim that is needed for the proof to follow.)

Let A be the original world and B the world with an added person, call her Barbara, whose life in B is at a wellbeing level m . Suppose that m lies in the neutral range. By the Intuition of Neutrality, this implies that B is neither better nor worse than A . We want to prove that B is not equally as good as A either. It will then follow that these two worlds must be incommensurable.

*Proof:*⁸ As shown above, there must be at least one level n in the neutral range that is either higher than m or lower than m . Now, consider a world C in which Barbara is added to the original world at wellbeing level n . Just as in B , no one else in C is affected by this addition.

Case 1: $n > m$. Since it is better for Barbara to live at a higher rather than a lower level of wellbeing, C is better for her than B . And it is equally as good as B for everyone else. We may therefore conclude that C is better than B . This is implied by the following general principle:

Suppose that worlds X and Y have the same population, I . (i) If X is better than Y for some individuals in I , while it is equally as good as Y for everyone else in I , then X is better than Y . (ii) If X and Y are equally as good for everyone in I , then X and Y are equally good.

This Pareto-like principle is an important part of the welfarist outlook. It establishes a minimal connection between personal and impersonal good. Following Broome (2004, section 8.2) we might call it the *Principle of Personal Good*.

This principle is compatible with the Intuition of Neutrality, even though the latter, on the standard interpretation of the neutral range, introduces a disparity

⁸ In its essentials, this proof is due to Broome; cf. his (2004), pp. 146ff.

between personal and impersonal good. While the Intuition of Neutrality focuses on value comparisons between worlds with partly different populations, the Principle of Personal Good – as stated above – is restricted to comparisons between worlds that have the same population. This restriction is crucial. Without it, the Principle of Personal Good would come into conflict with the Intuition of Neutrality. To see this, note that if $n > m$ and both m and n are in the neutral range that stretches upwards from zero, n must be a positive level of wellbeing. But then C , in which Barbara's level is n , is better for her than A : It is better for her to live a life at a positive wellbeing level than not to exist at all. At the same time C is equally as good as A for everyone else. Thus, in the absence of the restriction, the Principle of Personal Good would imply that C is better than A . However, since n lies in the neutral range, the Intuition of Neutrality implies that C is not better than A .

It is an interesting issue whether and how principles such as the Principle of Personal Good can be extended to comparisons between worlds with variable populations. The argument I have just given shows that the straightforward extension of this principle would be in conflict with the Intuition of Neutrality *if* the standard conception of the neutral range is assumed. This argument doesn't go through, however, if the neutral range is re-interpreted as will be done below, in section 3. Indeed, the straightforward extension of the Principle of Personal Good will then become possible without getting into conflict with the re-interpreted Intuition of Neutrality: The restriction to worlds with the same population will not be needed.

But let us return to the main proof. Suppose, for *reductio*, that B is equally as good as A . Then C , which by the Principle of Personal Good is better than B , must be better than A . (Betterness is transitive across equal goodness.) But since n just as m lies in the neutral range, the Intuition of Neutrality implies that C is *not* better than A . We must therefore conclude that B and A are not equally good.

Case 2: $n < m$. The Principle of Personal Good now implies that B is better than C . Consequently, if B were equally as good as A , A would also be better than C . But this again is excluded by the Intuition of Neutrality.

This completes the proof.

Thus, adding a person with a wellbeing level that lies in the neutral range creates an incommensurability. But how is this incommensurability to be explained? In the next section I suggest an explanation.

2. Analysis of value relations and neutral-range utilitarianism

I will start this section with rehearsing my general proposal as to how one might analyse different value relations, including the relation of incommensurability. (Cf.

Rabinowicz 2008, 2012.) I will then suggest, following Rabinowicz (2009), how this account can be applied to the problem at hand.

My account of value relations adheres to the general format of the fitting-attitudes analysis of value (FA-analysis). On this approach, value statements are interpreted as normative assessments of pro- and con-attitudes towards evaluated objects. As for statements of value relations, it is then natural to interpret them as normative assessments of preferences regarding the compared objects. Betterness and equal goodness are analyzed as follows:

x is *better* than y =_{df} x ought to be preferred to y .

x is *equally as good* as y =_{df} x and y ought to be equi-preferred.

Consequently, items x and y are incommensurable iff none of them ought to be preferred to the other nor ought they be equi-preferred.

There are two levels of normativity, the strong level of requirement ('ought') and the weak level of permission ('may'). Allowing for weak normativity as regards preferences makes room for further types of value relations. In particular, we can define the notion of parity, which is the typical (though not the only one) form of incommensurability:

x and y are *on a par* =_{df} x may be preferred to y and y may be preferred to x .

Thus, in cases of parity, opposing preferences regarding the compared items are permissible.

This approach to value relations can easily be formalized. In Rabinowicz (2008), I proposed the following *intersection modelling*:

The modelling has two components: the domain D of items that are compared and the class K of all permissible preference orderings of this domain.

Betterness is defined as required preference:

x is *better* than y iff x is preferred to y in every ordering in K .

This is just another way of saying that x ought to be preferred to y : every permissible preference ordering of the domain must incorporate this preference.

Analogously, equal goodness is defined as required equi-preference:

x and y are *equally good* iff they are equi-preferred in every K -ordering.

In case of parity, opposing preferences are allowed:

x and y are *on a par* iff x is preferred to y in some K -orderings and y is preferred to x in some other K -orderings.

Betterness is a transitive and asymmetric relation and equal goodness is transitive and symmetric. In addition, betterness is transitive across equal goodness. To guarantee that these conditions hold, we need to impose certain minimal formal constraints on the preference orderings in K . x will be said to be *weakly preferred* to y iff x is either preferred to y or equi-preferred with y . Indeed, we might just as well take this relation of weak preference to be our primitive concept and then define preference and equi-preference as, respectively, the asymmetric and the symmetric parts of weak preference: x is preferred to y iff x is weakly preferred to y but not vice versa; x is equi-preferred with y iff x is weakly preferred to y and vice versa. What we then need to assume about the orderings in K is that in each of them weak preference is reflexive and transitive. This constraint on permissible orderings is what gives us the desired formal properties of betterness and equal goodness.

Now, how can we apply this general modelling to the problem at hand? As the domain D we now take the set of possible worlds. If we accept the basic tenet of welfarism – the view that the value of a world is fully determined by the wellbeing levels of the individuals in that world – we can, for simplicity, identify each world with a wellbeing distribution: an assignment of (lifetime) wellbeing levels to the individuals that exist in this world. The question now is what the class K of permissible orderings of this domain is supposed to look like. Specifying K will determine the value relations that obtain between the possible worlds in D .

To specify K is thus to provide a substantive axiology: more precisely, a substantive population axiology. As is well-known, population ethics is haunted by conflicting intuitions and impossibility results. (Cf. Arrhenius 2000, 2011, 2016, and forthcoming. For a critical discussion, cf. Carlson, this volume.) It is a fair conjecture that no axiology can accommodate all plausible principles concerning value comparisons between worlds with variable populations. Here, I will focus on a utilitarian axiology, not because I find it fully satisfactory, but because it is a relatively plausible and at the same time a very simple form of welfarism. It is a simple and definite theory to work with. However, the kind of utilitarian axiology I want to consider has to make room for incommensurabilities in world comparisons, in order to accom-

moderate the Intuition of Neutrality. In Rabinowicz (2009) I called it *neutral-range utilitarianism*.

Let $w, w', \dots$ stand for lifetime wellbeing levels. We assume, until further notice, that these levels are measurable on a ratio scale: Thus, only the unit of measurement is arbitrarily chosen. (This strong measurability assumption will be relaxed in Section 3 and then much further relaxed in Section 4.)

Now, one way to think of the neutral range is that wellbeing levels in that range are potential candidates for being permissible ‘critical levels’ – permissible benchmarks – for preference. It is permissible to choose any of them, say level w , as the maximum below which adding new lives to the population is dispreferred. If a wellbeing level w lies in the neutral range, it can thus be used as a benchmark for determining the position of each world in the preference ordering. This position is obtained by summing up, for all individuals in the world in question, the surpluses and the deficits in their wellbeing, as compared with the chosen benchmark. Thus, let $I(A)$ be the set of individuals that exist in a world A and let w_{iA} be the wellbeing level of an individual i in A . The sum of wellbeing surpluses and shortfalls in A , relative to benchmark w , equals

$$\sum_{i \in I(A)} (w_{iA} - w).$$

The higher this sum is, the higher the world ends up in the preference ordering determined by w . Any two worlds for which this sum is the same occupy the same position in the ordering.

In this way, different preference orderings in K correspond to different choices of critical levels from the neutral range. A choice of a particular level specifies the point beyond which it is preferred to add a life to the population and below which such addition is dispreferred. Different choices of critical levels w from the neutral range (though only from that range) are all admissible and each such w induces a permissible preference ordering P_w on the set of worlds. Note that every P_w is a complete ordering: it contains no gaps. That is, for any two worlds, one of them is preferred in P_w to the other or they are equi-preferred in P_w . Class K might also include gappy preference orderings; it should be permissible to have incomplete preferences between worlds. I will return to this possibility in a moment. But if we disregard it, we can define the relations of betterness, equal goodness and parity between worlds as follows:

A world A is *better* than a world B iff for all w in the critical range, A is preferred to B in P_w .

A is equally as good as B iff for all w in the neutral range, A is equi-preferred with B in P_w .

A is on a par with B iff for some w and v in the neutral range, A is preferred to B in P_w and B is preferred to A in P_v .⁹

To illustrate, consider again the world B in which Barbara is added to the original world A at level m , where m belongs to the neutral range. Since this range contains more than one wellbeing level and since levels are linearly ordered from higher to lower, there must be some level n in that range that is higher or lower than m . Suppose that $n > m$. Then A is preferred to B in P_n but not in P_m . (In P_m , A and B are equi-preferred.) Analogously, if $n < m$, then B is preferred to A in P_n but not in P_m . Thus, in either case, K will contain preference orderings that differ from each other in how they rank A and B . This means that B is incommensurable with A , neither better nor worse than the latter, nor equally as good. Indeed, if the neutral range contains both levels higher than m and lower than m , A and B are on a par.

What about gappy preference orderings in K ? Let W be any non-empty subset of wellbeing levels in the neutral range. For example, W might be a sub-interval of that range. Let P_W be the set of complete preference orderings induced by the wellbeing levels in W . Plausibly, the intersection of P_W – the common part of the orderings in P_W – is also a permissible preference ordering of worlds.¹⁰ This intersection $\cap P_W$ will contain gaps if W contains more than one wellbeing level. Intuitively, $\cap P_W$ represents preferences of someone who is undecided between levels in W as potential benchmarks – someone who has not made up his mind as to where exactly to draw the line between preferred and dispreferred life additions.

Introducing gappy orderings into K in this way does not affect the extensions of the four typical value relations between worlds: better, worse, equally as good, and on a par. Nor does it affect the incommensurability relationships between worlds. Therefore, in most cases, there's no need to consider these incomplete K -orderings.

The population axiology described above is what I call *neutral-range utilitarianism* (NRU, for short). It combines total-sum utilitarianism with the Intuition of Neutrality. This axiology is formally identical with the theory that has been put forward by Blackorby, Bossert and Donaldson (1996). They call it the “incomplete critical-level utilitarianism” (ICLU). ICLU is a generalization of the more familiar

⁹ In typical cases of parity, there will also be some u such that A is equi-preferred with B in P_u . Indeed, this will hold in all mere-addition cases I here discuss. But in the interest of greater generality I abstain from imposing this condition as part of the definition of parity.

¹⁰ In this intersection of P_W , world A is weakly preferred to world B iff A is weakly preferred to B in every ordering in P_W .

critical-level utilitarianism (CLU), originally proposed by Blackorby and Donaldson (1984). CLU is a utilitarian theory that picks out a specific non-negative wellbeing level w as the unique critical level and then lets the value ordering of worlds coincide with P_w . Thus, on CLU, world A is better than world B iff $\sum_{i \in I(A)} (w_{iA} - w) > \sum_{i \in I(B)} (w_{iB} - w)$. If these two sums are equal, the two worlds are equally good. CLU entails that the value ordering of worlds is complete: there are no incomensurabilities. As this completeness of evaluation might well be questioned, ICLU has been proposed as a less categorical option. On ICLU, the value ordering of worlds is taken to be a compromise between different complete value orderings that are generated by different choices of critical levels. As such a compromise, it retains what is common to the alternative complete evaluations and leaves gaps at places where the complete evaluations disagree.

Clearly, even though ICLU and NRU are structurally identical theories, they differ in their philosophical motivations. On ICLU, the ultimate value ordering is a compromise between different complete value orderings, while on NRU it is instead the intersection of permissible *preference* orderings. NRU is based on the analysis of value relations in terms of permissible preferences. This anchoring in the FA-format of analysis is absent in ICLU. Indeed, the philosophical motivation for ICLU has never been made very clear by its proponents and it is therefore probably not accidental that in their subsequent publications Blackorby, Bossert and Donaldson have reverted to CLU as their favoured axiology.

NRU is also structurally identical to Broome's (2004) version of CLU. On this version, while there is just one critical level, its precise location is *indeterminate*. Instead of the neutral range we thus have a zone of indeterminacy – the set of wellbeing levels such that it is indeterminate which of them is the critical level but determinate that it is one of them. From this zone of indeterminacy, different precisifications of the theory choose different levels as the critical one and offer complete evaluations of possible worlds based on these choices. In the standard supervaluationist manner, what is common to all the precisifications (i.e., their intersection) is determinately true according to this axiology. Statements that hold on some precisifications but not on others are neither true nor false. They have an indeterminate truth value. Statements that don't hold on any precisification are false. On Broome's proposal, there are thus no incommensurabilities between worlds; instead, we have indeterminacies concerning their mutual value relations. Again, this axiology, in spite of its close structural similarity to NRU, has a different philosophical interpretation.

Let's move on. We have seen how NRU is supposed to work. But how plausible is it as a population axiology?

As is well known, the standard (total-sum) utilitarianism implies the Repugnant Conclusion (see Parfit 1991[1984], ch. 17):

Repugnant Conclusion: For any world whose inhabitants have excellent lives, there is a better possible world all whose inhabitants have lives barely worth living – lives that are good but only barely so.

Call the former and the latter world the Happy World and the Drab World, respectively. If the Drab World has a sufficiently large population, its total sum of wellbeing will exceed that of the Happy World. Increases in population size compensate and indeed outweigh losses in life quality.

Unlike total-sum utilitarianism, NRU does not allow for such facile compensations. Barely good lives, i.e. lives at very low positive levels of wellbeing, have wellbeing levels within the neutral range. On the standard conception of that range, it stretches from zero upwards to some relatively high (though not too high) level of wellbeing. Thus, adding lives with low positive levels of wellbeing to the world does not make it better. On NRU, the Repugnant Conclusion is avoided.

What is not avoided is the Weak Repugnant Conclusion – a claim that is just like the Repugnant Conclusion, but with “better” replaced by “not worse”. If the Drab World contains sufficiently many people, then, on NRU, it will be incommensurable with the Happy World. Indeed, the two worlds will be on a par. On some choices of a benchmark w from the neutral range, very close to zero (closer than the level of drab lives), the Drab World will be preferred to the Happy World, while choosing a higher w will yield the opposite preference. While Weak Repugnant Conclusion might be hard to accept, it is not as outrageous as the original Repugnant Conclusion.

Apart from the Repugnant Conclusion, NRU also avoids the so-called Sadistic Conclusion, which plagues CLU, the critical-level utilitarianism (see Arrhenius 2000):

Sadistic Conclusion: For any world whose inhabitants have terrible lives, there is a worse possible world whose inhabitants all have lives worth living (if only barely so).

CLU has this sadistic implication because the wellbeing levels of some lives worth living are lower than the critical level. Thus, on CLU, each such life detracts from the value of the world. Consequently, if the population of the Drab World is sufficiently large, this world will be worse than the Terrible World.

NRU avoids the Sadistic Conclusion if the neutral range goes all the way down to the zero level of wellbeing, as the standard interpretation of this range would have

it. Low positive wellbeing levels lie within the neutral range. Therefore, adding lives at such levels does not detract from the value of the world: it does not make the world worse.

But again, NRU does not avoid the Weak Sadistic Conclusion, i.e., the claim which is just like the Sadistic Conclusion, but with “worse” replaced by “not better”. The Drab World will not be worse than the Terrible World, but if its population is large enough, it will not be better. It will be incommensurable (on a par) with the Terrible World. It is a worrying and highly implausible implication.

This in itself might be a sufficient reason to consider another, alternative population axiology. But there is a more basic reason as well: As it stands, the Intuition of Neutrality is problematic.

3. The Intuition for Neutrality re-interpreted

As we have seen, on the standard conception of the neutral range, the Intuition of Neutrality introduces a disparity, a hiatus, between personal and impersonal goodness. A life might be good for a person who lives that life, better than non-existence, but still the addition of such a life to the world might not be impersonally good: It might not make the world better. It doesn’t make it better if the wellbeing level of the added life lies within the neutral range.

The disparity between what’s good for a person and what’s good for the world will be seen by some as an appealing feature of the Intuition of Neutrality, indeed, as the reason to adopt it in the first place. But others will consider it highly problematic; they will see it as foreign to the welfarist outlook.

I am now going to consider a re-interpreted version of the Intuition – one that removes the disparity. This new version of the Intuition is obtained by a re-interpretation of the neutral range: The idea is to identify it with the range of wellbeing levels at which a life is *neither good nor bad* for the person who lives or could live that life – neither better nor worse for her than non-existence. In other words, the *impersonal neutral range* – the range of wellbeing levels at which additions of lives make the world either better or worse – is now identified with the *personal neutral range* – the range of wellbeing levels at which a life is neutral for the person who lives (or could live) this life: levels at which this life is neither good nor bad for her. (Cf. Rabinowicz 2009, pp. 390f, for this suggestion. See also Gustafsson 2016, where this suggestion is adopted and elaborated.¹¹)

Note that this re-interpretation presupposes that such personal neutral range does exist, i.e., that there is more than one wellbeing level at which a life is personally

¹¹ Gustafsson uses a slightly different terminology, though, and, more importantly, he does not define the personal value of a life by comparing it to non-existence.

neutral. And, more specifically, it presupposes that personally neutral lives can be better or worse for us than other personally neutral lives, despite the fact that all such lives are neither better nor worse for us than non-existence. How is it possible? The answer should be obvious by now. It can only be possible if lives at these personally neutral levels are incommensurable in their personal value with non-existence.

The argument for this claim is analogous to the one I have presented in section 1 above, in connection with impersonal neutrality. Here, I give it in a simplified and condensed version. Thus, suppose that levels m and n are both personally neutral and $m < n$. By assumption, lives at these levels are neither better nor worse for the persons who live them than non-existence. Nor can any of these lives be equally as good for them as non-existence, as the following argument shows. One of these lives, the n -life, is personally better than the other, i.e., it is such that it would be better for a person to live this life than to live the other life. But this means that if the m -life were personally equally as good as non-existence, then the n -life would be personally better than non-existence, contrary to the assumption. (The argument here depends on personal betterness being transitive across personal equal goodness.) On the other hand, if the n -life were personally equally as good as non-existence, then the m -life would be personally worse than non-existence, again contrary to the assumption. Thus, both of these lives must be incommensurable in their personal value with non-existence. More generally, if every personally neutral life is better or worse than some other personally neutral life, as it must be if there is more than one wellbeing level in the personally neutral range and wellbeing levels are linearly ordered, then every such life must be incommensurable in its personal value with non-existence.

How is such incommensurability in personal value between a life and non-existence to be understood? To answer this question, I again need to appeal to the FA-format of analysis, but this time apply it to personal value. There is no clear consensus among FA-analysts as to how personal value – goodness for a person – should be understood. I will adopt the view that what is good for a person is what anyone who cares for her ought to wish or desire for her sake.¹² Admittedly, this suggestion is not very precise: both the notion of caring for someone and the notion

¹² A proposal roughly along these lines was put forward by Darwall (2002). For a competing FA-account, according to which what is good for a person is what *anyone* (and not just those who care for her) ought to favour for her sake, see Rønnow-Rasmussen (2011). (Cf also Rønnow-Rasmussen 2018.) The normative reach of the latter account is much wider than that of the former – too wide, I think. Cf. Taurek (1977, p. 304): “When I judge of two possible outcomes that the one would be worse (or better) for this person or this group, I do not, typically, thereby express a preference between these outcomes. Typically, I do not feel constrained to admit that I or anyone *should* prefer the one outcome to the other.” Though even Taurek would agree, I suppose, that I should have this preference if I care for the person or group in question.

of wishing/desiring something for someone's sake would need clarification.¹³ Hopefully, though, the main idea of this proposal is sufficiently clear. Extending it to personal value relations is straightforward:

x is better for i than y iff anyone who cares for i ought to prefer x to y for i 's sake.

x is equally as good for i as y iff anyone who cares for i ought to equi-prefer x and y for i 's sake.

Incommensurability between x and y obtains for i whenever neither of these items is better for i than the other, nor are they equally as good for i .

Parity for i between x and y – the most typical form of incommensurability in personal value – obtains when it is permissible for anyone who cares for i to prefer x to y for i 's sake and likewise permissible for them to have the opposite preference.

The idea is now that this kind of personal parity can obtain between a life and non-existence. For certain wellbeing levels – the ones in the personal neutral range – it is permissible, for the sake of a person who might live a life at this level, to prefer that she has this life rather than she does not exist and likewise permissible to have the opposite preference – permissible, that is, for anyone who cares for that person.¹⁴

Normally, in cases of parity, equi-preference with regard to the items on a par is also permissible. If it is permissible to prefer x to y and likewise permissible to have the opposite preference, then it should also be permissible to equi-prefer x and y . This also applies to cases in which a life is personally on a par with non-existence: it should be permissible, for the sake of a person who could live that life, not only to

¹³ Just to give an example of an issue that might need clarification: Can one care for an individual i and desire something for i 's sake if i does not exist (but *could* exist)? In principle, it should be possible: i in these locations appears in an intensional context; Neither caring for i nor desiring something for i 's sake is a relation in which i is one of the relata. (Just as, say, thinking of Pegasus is not a relation in which Pegasus is one of the relata.) But this would mean that something can be good for i or better for i even if i does not exist, contrary to what I have assumed above, in Section 1. Should we welcome this implication? It would make personal value comparisons between someone's life and her non-existence even easier than I have previously suggested. On the other hand, caring for someone or desiring something for her sake requires that one is at least able to identify that person, which is difficult in case of non-existents. So perhaps we should evade this problem and simply add to the analysis of what it means that something is good or better for i an extra condition that i does exist? This would be easy, but rather ad hoc.

¹⁴ For a suggestion as to why opposing preferences can be permissible in cases like this, see the next section, footnote 15.

prefer one of these alternatives to the other, but also to equi-prefer them – to be indifferent between her having this life and her non-existence. This point will be of some importance below.

Let me regiment the terminology. I will say that a life is personally better than (worse than, equally as good as) another life iff it is better (worse, equally as good) for a person to have the former life than (as) the latter. Analogously, a life is personally better than (worse than, equally as good as) non-existence iff it is better (worse, equally as good) for a person to have this life than (as) not to exist. In the same vein, I will talk about personal incommensurability, or parity, between a life and non-existence, or between one life and another life.

We have three kinds of lives:

A (personally) good life = a life that is personally better than non-existence.¹⁵

A (personally) bad life = a life that is personally worse than non-existence.

A (personally) neutral life = a life that is neither (personally) good nor bad.

The re-interpreted Intuition of Neutrality only applies to the (personally) neutral lives: According to the Intuition, adding such lives is impersonally neutral; it does not make the world either better or worse. This re-interpretation removes the disparity between personal and impersonal goodness. It is fully compatible with the re-interpreted Intuition that adding good lives (even such that are barely good) always makes the world better, just as adding bad lives always makes it worse. This also means that the re-interpreted Intuition of Neutrality is compatible with the Principle of Personal Good that is no longer restricted to the comparisons between worlds sharing the same population:

Unrestricted Principle of Personal Good:

Let $I(X)$ and $I(Y)$ be the populations of worlds X and Y , respectively, and let I be the union of $I(X)$ and $I(Y)$.

- (i) If X is better than Y for some individuals in I , while it is equally as good as Y for everyone else in I , then X is better than Y .

¹⁵ This is a context-independent notion of a good life. Ordinarily, when we say that someone's life is good, we implicitly compare it with typical lives that people have in a given social context. Here, though, I aim at a minimal standard of a life's goodness – one that does not vary with social circumstances.

- (ii) If X and Y are equally as good for everyone in I , then X and Y are equally good.

Until further notice I continue to assume that wellbeing levels are linearly ordered, from higher to lower. (This assumption will be given up in the next section, though.) We can thus think of the wellbeing levels as ordered along a vertical axis, with the personally neutral range located below the levels of good lives and above the levels of bad lives. I can no longer assume, though, that wellbeing is measured on a ratio scale: There is no longer a non-arbitrary zero level of wellbeing. A non-arbitrary zero level could be defined as the level of a life that is personally equally as good as non-existence. But, as we have seen, on the current picture no lives are like that. Good and bad lives are, respectively, personally better and worse than non-existence, while personally neutral lives are all personally incommensurable with non-existence. Thus, on this new picture, measurement of wellbeing on a ratio scale is no longer available. Still, we can continue to assume that wellbeing is cardinally measurable: Wellbeing levels can be represented on an interval scale, with only the unit and the zero point being arbitrary. Indeed, we have available a scale of wellbeing that is somewhat stronger than the interval scale but still weaker than the ratio scale: While the zero level of wellbeing is arbitrarily chosen, this arbitrariness has limits. It is reasonable to require that the zero level should be chosen from the personally neutral range. The reason is that for each wellbeing level in this range, and only for those levels, it is permissible to have preferences that equate that level with non-existence. To put it more precisely: If the life at such level is personally on a par with non-existence, then, as we have seen above, it should be permissible, for the sake of a person who might have this life, to be indifferent between her having this life and her non-existence.¹⁶ In the preference ordering that equates this life with non-existence, its level can therefore be taken as the zero point. In this sense, the levels of personally neutral lives, and only those levels, are permissible candidates for the zero point of the scale.

Indeed, each such permissible scale, with the zero point chosen from the personally neutral range, might be thought of as a *ratio scale* for preferential assessment of lives. This ratio scale specifies a particular permissible configuration of preference strengths for or against different lives, where these preferences are thought of as being held for the sake of a person who could have the lives in question. Thus,

¹⁶ Strictly speaking, each level in the personally neutral range will be on a par with non-existence only if this range forms an *open* interval. If this interval is closed, lives at the upper bound and at the lower bound of the range still are incommensurable with non-existence, but they aren't on a par with it. It is not permissible to prefer a life at the upper bound of the range to non-existence, nor to prefer non-existence to a life at the lower bound. However, even for each of these boundary levels it is permissible to have preferences that rank lives at this level equally with non-existence.

instead of a unique ratio scale of wellbeing, we now have a *set*, call it S , of permissible preferential ratio scales to work with.

This allows us to re-interpret and reformulate *neutral-range utilitarianism* (NRU). Consider any scale s in S . Let $I(A)$ stand for the set of individuals that exist in a world A and, for any individual i in I , let $s(i, A)$ be the measure of preference regarding i 's life in A , as measured on scale s . $s(i, A)$ specifies the extent to which i 's life in A is preferred or dispreferred, as the case may be, for i 's own sake, as compared with her non-existence. We can then determine the total sum of the s -values of all lives in A : $\sum_{i \in I(A)} s(i, A)$. The higher this sum is, the higher is the position of A in the preference ordering of worlds induced by s . The NRU-betterness relation on the set of possible worlds can now be defined in terms of these orderings: World A is better than world B iff A is ranked higher than B in all permissible preference orderings of worlds, i.e. in all orderings induced by the different scales in S .

What impact does the re-interpretation of the neutral range have on NRU's intuitive appeal? The Sadistic Conclusion is, of course, still avoided: Adding bad lives to the world always detracts from its utilitarian value. Indeed, the Weak Sadistic Conclusion is now avoided as well: Any world whose inhabitants have good lives is better than a world inhabited by people with bad lives, worth not living. But, on the other hand, on this re-interpretation re-instates the Repugnant Conclusion: On the re-interpreted account, good lives, even if they are only barely good, add to the value of the world and these additions do not diminish in value as more and more such lives are being added. Therefore, for any world in which everyone's life is excellent, there will be a better world, with a much larger population, in which everyone's life is still good but only barely so.

However, in Rabinowicz (2009) I pointed out that the repugnancy of this re-instated Repugnant Conclusion now is assuaged, if not altogether removed. What in my view made the original Repugnant Conclusion intuitively repugnant was the *short distance* between barely good lives and lives that are positively bad, if only barely so. The former were supposed to be only marginally better than the latter. But, on the re-interpreted conception of the neutral range, the distance between barely good and barely bad lives might well be quite considerable. As I put it: "Lives that are worth living, however modest, cannot be only marginally better than lives that are worth not living, i.e. that are worse than non-existence, if these two kinds of lives are separated by a personal neutral range of non-negligible size." (ibid., p. 406) Indeed, on this picture, barely good lives might be considerably better than drab lives of 'muzak-and-potatoes' variety that Parfit found so unappealing. It is plausible to think that such drab lives are personally neutral rather than positively worth living; plausibly, they are neither better nor worse for us than non-existence.

But then the Repugnant Conclusion does not seem to be so repugnant anymore.¹⁷

In the next section, however, we shall see that this reassuring diagnosis might have been premature.

Before concluding this section, I should mention that there is another problem with NRU, which I am not going to discuss in this paper. I have in mind what Broome (2004) calls the ‘greediness’ of neutrality. Suppose that the neutral range contains both level m and levels lower than $m-k$, for some $k > 0$. As is easily seen, NRU implies that adding to a world A a person with a life at level m , while at the same time decreasing by k units the wellbeing of one of the originally existing persons results in a world, B , that is incommensurable with A . For while choosing m as the benchmark gives rise to a preference ordering of worlds in which A is ranked above B , setting the benchmark at a wellbeing level lower than $m-k$ yields a preference ordering in which B is ranked above A .¹⁸ Thus, as Broome (2004, p. 170) puts it: “Incommensurateness [...] is a sort of greedy neutrality, which is capable of swallowing up badness and goodness and neutralizing it. This is implausible [...]”

This problem is exacerbated given my re-interpretation of the neutral range. On this re-interpretation, the life of the person added in B is not even good for her (though it is not bad for her either). But B still is not worse than A , even though it makes life positively worse for one of the originally existing people. This bad thing about B , as compared with A , is swallowed up by B ’s incommensurability with A . Effects like this might seem implausible, but I am inclined to believe that they should be accepted. They are just part of the package that comes with the existence of a neutrality range. For a critical discussion of the greediness objection, see Rabinowicz (2009).

4. Complicating the picture—incommensurable lives

Let us define a *strictly neutral* life as a life that is personally equally as good as non-existence. Obviously, all strictly neutral lives are neutral, i.e., neither personally better nor personally worse than non-existence, but the converse doesn’t hold. As we have seen in the preceding section, if a neutral life is personally better or worse

¹⁷ An essentially similar treatment of the Repugnant Conclusion is defended by Gustafsson (2016, who re-interprets the neutral range in the way as I have done in Rabinowicz (2009).

¹⁸ The same conclusion – that B is incommensurable with A – can also be established using weaker premises, if one proceeds less directly and considers some other possible worlds as well, along with A and B . It is then sufficient to rely on the utilitarian balancing of gains and losses only in comparisons between worlds that share the same population. Furthermore, there is then no need to assume my account of incommensurability, in terms of divergent permissible preference orderings. (Cf. Broome, 2004, p. 170, and Rabinowicz 2009.)

than some other neutral life, then it cannot be strictly neutral: The hypothesis that it is personally equally as good as non-existence cannot be upheld, on pain of a contradiction. We can refer to such lives as *weakly neutral*. A life is weakly neutral iff it is incommensurable with non-existence in its personal value.

If strictly neutral lives can exist, along with weakly neutral lives, we must give up the assumption that all lives' wellbeing levels are linearly ordered. The wellbeing level of a strictly neutral life cannot be fitted into such an ordering. Unlike weakly neutral lives, a strictly neutral life cannot be personally better or worse than any other neutral life. And it cannot be personally equally as good as any weakly neutral life: By the transitivity and symmetry of equal goodness, it can only be equally as good as other lives that are equally as good as non-existence, i.e., it can only have this relation to other strictly neutral lives.

Postulating the possibility of strictly neutral lives along with lives that are weakly neutral thus requires that we allow for the existence of lives that are mutually incommensurable in their personal value. But, on reflection, this is something that we should allow for anyway. Surely, it is implausible to insist that for any two lives with different wellbeing levels, the wellbeing level of one must be higher or lower than that of the other life. Life wellbeing is a *many-dimensional* concept: Specifying its level requires characterizing a life with respect to several relevant dimensions. One life might be better than another in some respects, and worse in other respects. At the same time, different weight assignments to the relevant respects of comparison might be permissible and the all-things considered preference ordering of lives will depend on how these respects are weighed against each other. Consequently, a life L might be preferred to another life L' , for the sake of a person who might live one of these lives, given one permissible weight assignment, and dispreferred given another. This would imply that L and L' are incommensurable in their personal value, or – what amounts to the same – that L and L' have different wellbeing levels even though none of these levels is higher than the other. Thus, we can no longer assume that wellbeing levels of lives are linearly ordered.¹⁹

Nevertheless, while we should for this reason be willing to accept that lives might well be mutually incommensurable in personal value, we might still wonder whether to allow for the existence of strictly neutral lives. Gustafsson (2016, section 5) is unwilling to admit this possibility. He uses a different label for lives of this

¹⁹ That some lives (namely, the ones that are weakly neutral) can be incommensurable with non-existence in their personal value has a similar explanation, by the way. A life typically has both desirable and undesirable components and its assessment on balance, in comparison with non-existence, might depend on how these components are weighed against each other. Since different weight assignments will normally be permissible, they might give rise to opposing permissible preferences regarding the life in questions: It might be permissible, for the sake of the person who could live this life, to prefer it, all things considered, to her non-existence, but it might also be permissible to have the opposite preference, for her sake.

kind²⁰ and he does not define them by comparing their personal value with non-existence – indeed, he questions the possibility of such comparisons. But he takes such lives, *if* they exist, to satisfy the following three principles:

For any strictly neutral life L and for every life L' ,

(i) L' is good iff it is (personally) better than L ,

(ii) L' is bad iff it is (personally) worse than L ,

(iii) L' strictly neutral iff it is (personally) equally as good as L .²¹

This squares with the definition I have proposed: As is easily seen, the definition of a strictly neutral life as one that is personally equally as good as non-existence entails (i), (ii) and (iii), assuming that the relation of equal personal goodness is transitive and that personal betterness is transitive across this relation.

One potential candidate for a strictly neutral life that Gustafsson (2016) considers, but finally rejects, is a life devoid of any good or bad components.²² An example of a life of this kind would, I take it, be a life in a permanent state of unconsciousness.²³ Gustafsson denies, however, that such a life is possible for a person. One is not a person if one is never conscious. Being a person requires having some psychological features and being in some psychological states, but no such states and features can be present if consciousness is permanently absent.

I don't think this shows what it is supposed to show. It doesn't show that a person could not live such a life. It is true that I wouldn't be a person if my whole life were spent in a coma. But it still is true that I – a person – could have had such a life (in which I wouldn't be a person). It is a possible life for me, a person. While it is a popular view that being a person is an essential property of persons, I think this view is mistaken: A person could have had a life in which she wouldn't be a person. But

²⁰ He simply calls them "neutral" and refers to lives I call weakly neutral as "blank" (or "undistinguished", in later versions of his draft). In my comments on Gustafsson, I will however, continue to use my own terminology, to avoid confusion.

²¹ Note that if a life L is weakly neutral, it need not satisfy (i) and (ii): a life L' might be better than L without being good and it might be worse than L without being bad. Rather than good or bad, respectively, L' might itself be weakly neutral.

²² Indeed, if such a life is to be strictly neutral, it should also lack weakly neutral components, i.e., it should lack components that it is permissible to prefer (for the sake of the person who could lead this life), but also permissible to disprefer.

²³ In Broome's terminology, a life without any good or bad experiences is a 'blank life', and a life spent in a coma is an example of a blank life. (Cf. Broome 2004, pp. 208f.)

then for me, a person, my counterfactual life as a non-person, in a permanent state of unconsciousness, might be equal in value with my (equally counterfactual) non-existence. Consequently, this life might be strictly neutral.

Whether a life in a coma in fact is strictly neutral might of course be questioned. On some views, such a life would be worse for me than non-existence, due to considerations relating to human dignity. Perhaps a better candidate for a strictly neutral life could be found. Or perhaps not. The claim that strictly neutral lives could exist is logically coherent, but it is not clear to me whether such lives really are possible. Perhaps it might be argued that, for any possible life, it is at least permissible to prefer it or permissible to disprefer it (or both) to non-existence, for the sake of a person who could live that life. In such a case, no possible life would be strictly neutral. Let me now, however, consider what it would mean if strictly neutral lives could exist.

A neutral-range utilitarian might find the possibility of such lives quite worrying, as it implies, as we have seen, that lives might be mutually incommensurable in personal value. One might think that such incommensurabilities would make NRU meaningless: It would no longer be possible to determine the total value of a world's population for different choices of benchmarks from the neutral-range.²⁴

But, as I have pointed out, even if there were no strictly neutral lives to reckon with, we would still have to accept that some lives can be mutually incommensurable in personal value. If it were true that such incommensurabilities make NRU meaningless, then this theory would not be worth serious consideration. I think, though, that this worry can be put to rest: NRU can accommodate incommensurabilities between lives. Let me explain how it can be done.

On the interpretation of NRU I have given in the preceding section, this theory assumes that different permissible preference orderings of worlds are generated by different permissible preferential ratio scales for assessing lives – scales whose zero points are drawn from the neutral range. If life wellbeing is cardinally measurable, as we have previously assumed, then these scales only differ in their choices of zero from the neutral range, i.e., in their choices of the neutral wellbeing level such that lives on that level are equi-preferred with non-existence on a given scale. (The scales might also differ in their choices of the unit of measurement, but this choice doesn't matter when we do utilitarian calculation in order to compare worlds with each

²⁴ This is indeed the reason why Gustafsson (2016), who proposes an ethical theory that is very much like NRU (he calls it 'critical-range utilitarianism') and who like myself works with the personalized neutral range, is so opposed to strictly neutral lives. However, the reasons he offers for rejecting such lives change in the later versions of his paper. There, the meaninglessness worry no longer is mentioned. And even in the original version he eventually suggests a version of critical-range utilitarianism that does allow for incommensurable lives. On that theory, utilitarian aggregation of wellbeing in a population is done at the level of life moments instead of the whole lives. But he still assumes that, at the level of life moments, there are no incommensurabilities.

other. And since world comparisons are done independently for each scale, it doesn't matter either if different scales use different units of measurement.) However, if incommensurabilities between lives are allowed, the assumption of cardinal measurability of wellbeing must be given up.²⁵ Not only will the choice of zero differ, as before, between different permissible preferential ratio scales, but now some such scales will also differ in their *ordering* of lives: They will differ in how they rank incommensurable lives against each other. For example, if lives L and L' are personally on a par, then some permissible scales will rank L above L' , while others will rank it below.²⁶ The set S of permissible ratio scales will thus be much more varied than it was previously assumed. Class K of permissible preference orderings of worlds will however still consist of the orderings that are generated by the scales in S , as before. For any scale s in S , the position of a world A in the world ordering P_s induced by s is determined by the sum $\sum_{i \in I(A)} s(i, A)$. The higher this sum is, the higher is A 's position in P_s . The value relations between worlds are then determined by this class K of permissible preference orderings of worlds, in the standard way: world A is better than world B iff it is ranked above B in all permissible orderings; A is equally as good as B iff it is equal-ranked with B in every permissible ordering; A is incommensurable with B iff none of these worlds is better than the other nor are they equally as good.²⁷

Clearly, on this version of NRU, there will be many incommensurabilities between worlds, many more than if all lives were commensurable in their personal value. But this does not, by itself, undermine a utilitarian axiology.

Nor does this account undermine the basic welfarist principle that all (impersonal) value comparisons between worlds are determined by the wellbeing levels of the individuals existing in those worlds. However, the wellbeing level of a life no longer can be represented by a single number. Instead, we can now represent it as a function that assigns a numerical value to every scale in S . These values specify how highly a life at this wellbeing level is assessed on each scale. If all these values are positive, the wellbeing level is positive; if they are all negative, the wellbeing level is

²⁵ What is not given up is cardinal measurability of the *strength of preference* for different lives, according to each permissible preferential scale. But a life's wellbeing, which is characterizable by the position of this life on all the different permissible preferential scales, is no longer cardinally measurable. Nevertheless, lives' wellbeing levels can still be given a numerical representation, as will be shown below.

²⁶ Note, though, that if strictly neutral lives can exist, they will be placed in all those scales at the zero level. On any such scale, the zero level is the level of lives that are equi-preferred, for the sake of a person who could have that life, with her non-existence. Thus, in one way or another, each scale will commensurate strictly neutral lives with lives that are weakly neutral.

²⁷ Strictly speaking, class K will also contain incomplete world orderings. But if any such permissible incomplete ordering is the intersection of some set of permissible complete orderings, then adding incomplete orderings to K will not affect the extensions of the relevant value relations between worlds: the relations of betterness, equal goodness, incommensurability, and parity.

negative. If neither holds, the wellbeing level is neutral. In particular, for a strictly neutral life, all the values are zero. Two lives have the same wellbeing level if they are characterized by the same function of this kind. One wellbeing level, w , is higher than another, w' , iff for all scales s in S , the s -value of the function that represents w is higher than the s -value of the function that represents w' . In other words, the wellbeing level of one life, L , is higher than that of another life, L' , iff on every scale in S , L is ranked above L' . We can also define what it means that one wellbeing level is at least as high as the other: This is the case iff for all scales s in S , the s -value for the former level is at least as high as the s -value for the latter level. (Note that, given this definition, a life might be at a wellbeing level that is at least as high as another life without the level of the former life being higher than or equal to that of the latter life.) As is easy to see, wellbeing levels are partially ordered by this at-least-as-high-as relation. In other words, the relation in question is reflexive, transitive and anti-symmetric. (Anti-symmetry means that if level w is at least as high as level w' and w' is at least as high as w , then w is identical to w' .)

Thus, the worry that NRU becomes meaningless if strictly neutral lives are allowed and, more generally, if lives are allowed to be incommensurable, is unjustified. It is still possible to define lives' wellbeing levels and carry out utilitarian calculations in order to determine the value of each world. This value is a function that for each scale s in S specifies the total sum of the s -values of the lives of the individuals that exist in the world in question.

It is another issue, however, whether NRU, so interpreted, is an intuitively appealing theory. This might be questioned. To explain why, let me first point out that allowing for strictly neutral lives leads to rather unexpected and unwelcome consequences. Suppose, for definiteness, that a life L , which is wholly spent in a coma, is strictly neutral, and consider a life L^+ that is slightly personally better than L . L^+ might, for example, be a life that also is mainly spent in a coma, apart from a very short period during which its subject is conscious and experiences a moderate sensory pleasure (and nothing else). Since L^+ is slightly (personally) better than L , which is (personally) equally as good as non-existence, it follows that L^+ is slightly (personally) better than non-existence. Which implies that L^+ is a good life, though only barely so. In the same way, we can think of a barely bad life, L^- , which is slightly (personally) worse than L and thus slightly (personally) worse than non-existence. L^- might be a life mainly spent in a coma, apart from a very short period during which its subject is conscious and experiences a moderate pain (and nothing else).

Now, note that the value distance between L^+ and L^- is short: L^+ is only marginally better than L^- . Thus, we now have to accept that there are good lives that are marginally better than bad lives.

But how can this be? Didn't we previously show that wellbeing levels of good lives

are separated from the wellbeing levels of bad lives by the neutral range, which might be quite extended?

That was then, though, before we gave up the assumption that wellbeing levels are linearly ordered. On the new picture, things look very differently. Indeed, as we already know, a strictly neutral life L , if such a life can exist, is incommensurable in its personal value with all lives that are weakly neutral. And it is arguable that the same applies to the barely good life L^+ and the barely bad life L^- . As has been noted by Joseph Raz, it is a “mark of incommensurability” that if this relation obtains between two items, then a small improvement or a small worsening of one of the items need not (and typically will not) remove their incommensurability (see Raz 1986, p. 26). Thus, since incommensurability obtains between the strictly neutral life L and all lives that are weakly neutral, it might well still obtain when L is replaced by L^+ or L^- .

This means that on the new picture a good life such as L^+ , and a bad life such as L^- might be incommensurable with a neutral life – a life that is neither bad nor good. Indeed, they might both be incommensurable with all neutral lives that are weakly neutral. Some good lives (such as L^+) need no longer be better than all lives that aren’t good and some bad lives (such as L^-) need no longer be worse than all lives that aren’t bad. This might be surprising, but it is an implication that we now must accept.

Note also that this implication does not strictly speaking require the possibility of strictly neutral lives. What it does require is the possibility of lives *close to strict neutrality*: lives that are only slightly better or slightly worse than non-existence. Since weakly neutral lives are incommensurable with non-existence, it is to be expected that they will also be incommensurable with lives such as L^+ and L^- .

Thus, the value distance between a barely good life and a barely bad life might be very short. The former might be only marginally better than the latter.²⁸ L^+ and L^- provide a case in point.²⁹ But this means that the Repugnant Conclusion which follows from the re-interpreted NRU regains its original repugnance. Just as on the standard total-sum utilitarianism, we are now driven to the conclusion that every

²⁸ In private communication, John Broome has suggested the a situation like this will arise if “there are pairs of lives such that one is better than non-existence and one is worse than non-existence, and they are very similar to each other.”

²⁹ But if we no longer assume that lifetime wellbeing is cardinally measurable, how can we say that L^+ is only *marginally* better than L^- ? It might be objected that in the absence of cardinal measurement, differences in lifetime wellbeing cannot be judged to be small or large. I don’t think, though, that this objection is compelling. Pockets of cardinality – areas in which size estimates of wellbeing differences are meaningful – might still exist, even though cardinal measurement no longer is possible in all cases – in all comparisons of lives’ wellbeing levels. Those are the areas in which all the permissible ratio scales in S agree in their assessments of differences between life levels. The difference between two wellbeing levels is small if it is small on each of the permissible scales.

world in which everyone has an excellent life is worse than some world in which everyone's life is only marginally better than a life worth not living.

This is bad news for neutral-range utilitarianism. But its adherents might be well-advised to stand fast and hold on to their view. The truly repugnant Repugnant Conclusion crucially depends on the possibility of lives close to strict neutrality. But this possibility, just as the possibility of strictly neutral lives themselves, has *not* been positively established. Until it has been done, which might never happen, NRU remains unchallenged. Its adherents might persist in denying that strictly neutral lives and lives close to strict neutrality really are possible.³⁰

However, just as a thought experiment, suppose someone proves to us that such lives indeed are possible. Then NRU will have to be given up unless we are willing to accept the repugnant Repugnant Conclusion. But, if we give up NRU, are there some claims from the preceding discussion that we still can we retain?

I think we can retain the Intuition of Neutrality itself, in its re-interpreted, personalized version:

Adding personally neutral lives to the world is impersonally neutral: It does not make the world either better or worse.

We can also retain the insight that:

Lives can be incommensurable with each other in their personal value.

In other words, wellbeing levels are not linearly ordered.³¹

Despite this acceptance of incommensurabilities between lives, we can, if we wish, hold on to some of the central tenets of welfarism:

(i) *The Unrestricted Principle of Personal Good,*

and the assumption that:

³⁰ A mere logical possibility of such lives is not enough to undermine NRU. I am indebted to John Broome for a discussion of this point (in private communication).

³¹ This of course also applies to levels in the neutral range, and it would apply to them even if there were no strictly neutral lives. Lives that are weakly neutral can be mutually incommensurable. But still, it is reasonable to suppose that for every such level *m* there is some neutral level that is higher or lower than *m*. This means that we can continue to uphold the assumption of the proof provided in Section 1. As we remember, that proof was meant to establish that a world in which a person is added at some such level *m* must be incommensurable with the original world.

(ii) *Value comparisons between worlds are determined by the wellbeing levels of individuals who exist in those worlds.*³²

Furthermore, if lives close to strict neutrality are possible, we can draw the rather surprising lesson that:

A good life need not be better than a neutral life and a bad life need not be worse than a neutral life.

A life that is better (worse) than non-existence is good (bad) but it need not be better (worse) than a life that is incommensurable with non-existence. It might be incommensurable with such a life.

This lesson generalizes: It is potentially applicable to all analyses of ‘good’ and of other monadic value predicates in terms of value comparisons with some *standard*. In this paper, when defining monadic value predicates, the items we have targeted have been lives and the standard we have chosen has been non-existence. But we can consider other items and choose other standards. Thus, suppose we consider some domain of items, pick out a standard σ and adopt the following definitions: For all items x in this domain,

x is good iff x is better than σ ,

x is bad iff x is worse than σ ,

x is neutral iff x is neither better nor worse than σ ,

x is strictly neutral iff x is equally as good as σ .

Then it does not follow that a good x must, by logical or analytical necessity, be better than a neutral y , or that a bad x must be worse than such a y .³³ These seemingly very plausible entailments do not obtain if y , while neutral, is not strictly so, i.e. if y is incommensurable with the adopted standard. In such cases, a good x will typically not be better than y if x is only slightly better than the standard. Likewise, a bad x

³² But, in view of the incommensurabilities between lives, wellbeing levels no longer can be assumed to be representable by single numbers. Their numerical representation has to be more complicated, as we have seen above.

³³ This has already been noted in Gustafsson (2016).

will typically not be worse than such y if x is only slightly worse than the standard.

Thus, some of the things we have learned have implications that go beyond population axiology.³⁴

References

- Arrhenius, G. (2000) "An Impossibility Theorem for Welfarist Axiology," *Economics and Philosophy* 16: 247–266.
- Arrhenius, G. (2011) "The Impossibility of a Satisfactory Population Ethics", in E. Dzhafarov and L. Perry (eds.), *Descriptive and Normative Approaches to Human Behavior*, World Scientific: 51–66.
- Arrhenius, G. (2016) "Population Ethics and Different-Number-Based Imprecision", *Theoria* 82: 166–181.
- Arrhenius, G. forthcoming, *Population Ethics: The Challenge of Future Generations*, Oxford University Press.
- Arrhenius, G., & Rabinowicz, W. (2010) "Better to Be Than Not to Be?" in H. Joas and B. Klein (eds.), *The Benefit of Broad Horizons*, Brill: Leiden, 399–421.
- Arrhenius, G., & Rabinowicz, W. (2015) "The Value of Existence," in I. Hirose and A. Reisner (eds.), *The Oxford Handbook of Value Theory*, Oxford University Press: Oxford, 424–444.
- Blackorby, C., Bossert, W., & Donaldson D. J., (1996), "Quasi-orderings and Population Ethics," *Social Choice and Welfare* 13: 129–150.
- Blackorby, C. & Donaldson, D. (1984), "Social criteria for evaluating population change". *Journal of Public Economics* 25:13–33.
- Broome, J. (1999) *Ethics out of Economics*, Cambridge University Press: Cambridge.
- Broome, J. (2004) *Weighing Lives*, Oxford University Press: Oxford.
- Broome, J. (2009) "Reply to Rabinowicz," *Philosophical issues* 19: 412–417.

³⁴ Early versions of this paper were presented at a conference on formal ethics in York in June 2017 and at a conference on philosophy and economics in Canberra the same year. I am indebted to the participants of these events for several useful suggestions. I also want to thank John Broome, Johan Gustafsson and Gustaf Arrhenius for their helpful written comments and discussion. Broome's challenging comments, in particular, have caused me to modify some of the conclusions of this paper.

- Bykvist, K. (2007) "The Benefits of Coming into Existence," *Philosophical Studies* 135: 335–62.
- Bykvist, K. (2015) "Being and Well-Being," in I. Hirose and A. Reisner (eds.), *Weighing and Reasoning*, Oxford University Press: Oxford.
- Carlson, E., this volume, "On Some impossibility Theorems in Population Ethics".
- Darwall, S. (2002), *Welfare and Rational Care*, Princeton University Press: Princeton and Oxford.
- Gustafsson, J. (2016), "Population Axiology and the Possibility of a Fourth Category of Absolute Value," draft. A revised version has appeared in *Economics and Philosophy* 36, 2020, pp. 81–110.
- Holtug, N. (2001) "On the Value of Coming into Existence," *Journal of Ethics* 5: 361–84.
- Johansson, J. (2010) "Being and Betterness," *Utilitas* 22: 285–302.
- Narveson, J. (1973) "Moral Problems of Population," *The Monist* 57: 62–86.
- Parfit, D. (1991 [1984]) *Reasons and Persons*, Clarendon Press: Oxford.
- Rabinowicz, W. (2008), "Value Relations," *Theoria* 74: 18–49.
- Rabinowicz, W. (2009) "Broome and the Intuition of Neutrality," *Philosophical Issues* 19: 389–411.
- Rabinowicz, W. (2012), "Value Relations Revisited," *Economics and Philosophy* 28: 133–164.
- Raz, J. (1986), *The Morality of Freedom*, Oxford: Clarendon Press.
- Rønnow-Rasmussen, T. (2011), *Personal Value*, Oxford University Press: Oxford.
- Rønnow-Rasmussen, T. (2018), "Fitting-Attitude Analysis and the Logical Consequence Argument," *The Philosophical Quarterly* 272: 560–579.
- Taurek, J. (1977), "Should the Numbers Count?" *Philosophy and Public Affairs* 6: 293–316.