philosophies

MDPI

Article

Axiological Retributivism and the Desert Neutrality Paradox

Tim Campbell

Institute for Futures Studies, 101 31 Stockholm, Sweden; timothy.campbell@iffs.se

Abstract: According to axiological retributivism, people can deserve what is bad for them and an outcome in which someone gets what she deserves, even if it is bad for her, can thereby have intrinsic positive value. A question seldom asked is how axiological retributivism should deal with comparisons of outcomes that differ with respect to the number and identities of deserving agents. Attempting to answer this question exposes a problem for axiological retributivism that parallels a well-known problem in population axiology introduced by John Broome. The problem for axiological retributivism is that it supports the existence of a range of negative wellbeing levels such that if a deserving person comes into existence at any of these levels, the resulting outcome is neither better nor worse with respect to desert. However, the existence of such a range is inconsistent with a set of very plausible axiological claims. I call this the desert neutrality paradox. After introducing the paradox, I consider several possible responses to it. I suggest that one reasonable response, though perhaps not the only one, is to reject axiological retributivism.

Keywords: retributivism; axiology; desert; intuition of neutrality; desert-neutrality paradox; population ethics; John Broome; Shelly Kagan; Larry Temkin

Citation: Campbell, T. Axiological Retributivism and the Desert Neutrality Paradox. *Philosophies* **2022**, *7*, 80. https://doi.org/10.3390/philosophies7040080

Academic Editor: Stephen Kershnar

Received: 4 May 2022
Accepted: 17 June 2022
Published: 15 July 2022

1. Introduction

According to axiological retributivism, people can deserve what is bad for them, provided that they are sufficiently evil, and an outcome in which some such evil person gets what she deserves, even if what she deserves is bad for her, can thereby have intrinsic positive value [1–3].

Larry Temkin proposes the following test of whether one is sympathetic to axiological retributivism.[1] Imagine that Adolf Hitler, perhaps the evilest agent in history, having intentionally caused the immense suffering of millions of innocent people, without a shred of remorse and with great personal satisfaction, is currently enjoying an after-life filled with the highest forms of physical pleasure, desire fulfilment, achievement, respect, self-actualization, and other personal goods. Next, imagine that during his glorious afterlife, Hitler comes down with a cold, which causes him to suffer a little bit and frustrates some of his desires. Temkin's reaction to this thought experiment is to think "Hooray for the cold!" In other words, he is inclined to believe that in this scenario Hitler suffering is a good thing. His suffering contributes intrinsic positive value to the outcome in which he exists. If, like Temkin, you think "Hooray for the cold!" then you are sympathetic to axiological retributivism. You might be sympathetic to axiological retributivism even if you do not have this reaction. For instance, you might think that Hitler gets what he deserves only if his suffering is intentionally inflicted on him by human agents as a form of just punishment. But if you are inclined to believe that Hitler, or some suitably evil agent, suffering in some circumstance would be intrinsically good in some respect, then you have an intuition that supports axiological retributivism.

In this paper, I raise a problem for axiological retributivism. The problem arises when one asks what this view implies about pairs of outcomes that differ with respect to the number and identities of deserving agents. It seems to me that most existing discussions of retributivism ignore this question, focusing narrowly on same-people cases (i.e., cases in which all and only the same people exist in the different outcomes that we are comparing

with respect to desert value). Important exceptions to this custom of focusing on same-people cases include Feldman [4,5] and Arrhenius [6,7]. While it may be useful, for the sake of simplicity, to limit discussion of retributivism to same-people cases, there is no good reason to think that retributivism applies only in such cases.

Unfortunately, when one tries to extend the application of axiological retributivism beyond same-people cases, it leads to a paradox that mirrors a familiar paradox in population ethics. The latter paradox arises when one tries to formulate a plausible population axiology (i.e., a theory of the goodness of outcomes, where these outcomes may differ with respect to the number, identities, and wellbeing levels of the people who exist in them) that can accommodate what John Broome calls the intuition of neutrality [8,9]. In its axiological interpretation, the intuition is that there is a range of positive wellbeing levels such that if a new person comes into existence at any of these levels, then this does not, by itself, make the resulting outcome either better or worse [8] (pp. 145–146). As Broome demonstrates in his book *Weighing Lives*, it seems impossible to accommodate this intuition without rejecting some seemingly obvious axiological claim, such as the claim that the relation 'better than' is transitive (i.e., for any outcomes x, y, z, if x is better than y, and y is better than z, then x is better than z) [8] (pp. 145–149, 180–183). I refer to this tension between the intuition of neutrality and other seemingly obvious axiological claims as the neutrality paradox. I argue that when axiological retributivists extend the application of their theory beyond same-people cases, they will be pressed to acknowledge a range of wellbeing levels such that if a new deserving person comes into existence at any of these levels, the resulting outcome is neither better nor worse with respect to desert, other things being equal. I argue that this claim, like the intuition of neutrality, seems impossible to accept without rejecting some seemingly obvious axiological claim. I call this the desert-neutrality paradox. I suggest that the desert-neutrality paradox poses a bigger problem for axiological retributivism than the neutrality paradox poses for population axiology. It would be unreasonable to respond to the neutrality paradox by rejecting population axiology altogether, but it would not be unreasonable to respond to the desert-neutrality paradox by rejecting axiological retributivism.

In Section 2, I outline what I take to be some core commitments of axiological retributivism, and I illustrate these using a tool introduced by Shelly Kagan—the desert graph [2] (chapter 3). In Section 3, I explain the neutrality paradox. In Section 4, I introduce the desert-neutrality paradox, and I consider its implications for axiological retributivism. I conclude in Section 5 by canvasing various ways out of the paradox. I suggest that one reasonable way out is to reject axiological retributivism.

2. Commitments of Axiological Retributivism

There are different possible views about what kinds of personal bads an evil person deserves. I will use the term 'suffering' to refer broadly to whatever the relevant personal bads are (pain, shame, humiliation, failure, constrained agency, etc.). There are also different possible explanations of why deserved suffering has intrinsic value, and of the kind of intrinsic value that it has [1]. I shall refrain from taking a stand on either issue.

In thinking about how to characterize axiological retributivism, I find it useful to focus on Temkin's example of Hitler and the cold. This example is useful partly since it illustrates the idea that a person can get more or less of what he deserves, or, to put it another way, that he can be closer or further away from what he *perfectly* deserves. If you think Hitler deserves to suffer, you probably think he deserves a great deal of suffering, not just the mild suffering of a cold. Catching cold might give Hitler some of the suffering he deserves, but he deserves more. In Temkin's thought experiment, Hitler enjoys a life with very little suffering and a great deal of what is good for him. He drastically underpays for his crimes. Retributivists will view such an outcome as being, in some respect, intrinsically bad.

Retributivists will also want to say that there is some amount of suffering that would be too much even for Hitler. It is bad if Hitler drastically underpays for his crimes, but it is also bad if he drastically overpays. This suggests that there is an amount of suffering that

Hitler perfectly deserves—an amount such that his getting more or less than that amount would make things worse with respect to desert than his getting exactly that amount [2] (p. 77).[2] Thus, retributivism seems to include at least following commitments:

1. The following are possibilities:
   (a) There exist people who are so evil that they perfectly deserve some amount of suffering.
   (b) It is true of some of those referred to in (a) that their having the amount of suffering that they perfectly deserve constitutes one respect in which the outcome in which they exist is intrinsically good.
2. For any person S, it is true that
   (a) S having an amount of suffering that is closer to the amount that S perfectly deserves constitutes one respect in which the outcome in which S exists is better, intrinsically, than an outcome in which the amount of S's suffering (if S suffers at all) is further from the amount that S perfectly deserves.
   (b) S having an amount of suffering that is sufficiently far from the amount that S perfectly deserves constitutes one respect in which the outcome in which S exists is intrinsically bad.

Everyone should recognize 1(a) as a core commitment of axiological retributivism. Commitment 2(b) captures the intuition that things are intrinsically bad in some respect if Hitler drastically underpays or overpays for his crimes. Commitment 2(a) captures the intuition that things are in some respect better if the amount of Hitler's suffering is closer to what he perfectly deserves. This commitment seems to explain why retributivists like Temkin cheer for the cold. Although the mild suffering of a cold may be just a drop in the bucket compared to what Hitler perfectly deserves, it is at least something. It would be worse if Hitler got none of the suffering he deserves. Axiological retributivism, as I understand it, implies that there is intrinsic positive value—intrinsic goodness—in a person having the suffering he perfectly deserves. Thus, I consider 1(b) a central commitment of this view. However, Shelly Kagan characterizes axiological retributivism somewhat more broadly than I do. According to Kagan, one who claims that desert has axiological relevance need not claim that people getting what they perfectly deserve has intrinsic positive value [2] (pp. 147–148). One could instead claim that people getting what they perfectly deserve is intrinsically neutral (neither good nor bad), and that people getting more or less than what they perfectly deserve has only intrinsic negative value. This view is not a version of what I am calling axiological retributivism, but it may avoid the desert-neutrality paradox and may therefore be attractive as a fallback position.

One can illustrate the commitments listed above using a tool first introduced by Kagan—the desert graph [2] (chapter 3). Following Kagan, I will assume that what people deserve is a certain level of wellbeing, which could be positive, or, if the person is sufficiently evil, negative. (I shall therefore understand 'suffering' in terms of having negative wellbeing.) It is now possible to plot desert value as a function of a person's wellbeing, while holding fixed that person's desert profile (i.e., everything about the person that determines exactly what level of wellbeing she deserves). A person's desert profile could include, for example, the magnitude of the wrongs she commits and her degree of culpability in committing these wrongs. For any person with a given desert profile, the function from different wellbeing levels she might occupy to different desert values is her desert function. This tells us what her having a certain wellbeing contributes to an outcome's desert value, other things being equal. The desert graph in Figure 1 represents a possible desert function for Hitler that captures the abovementioned commitments of axiological retributivism [2] (p. 76).

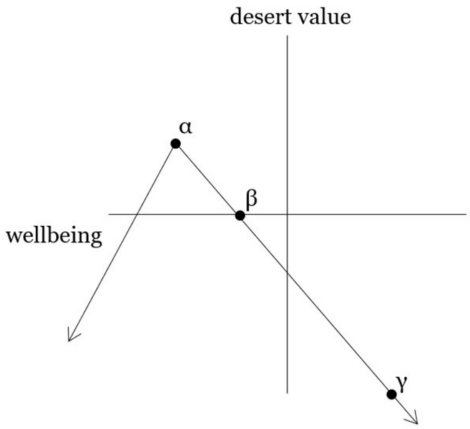

Figure 1. Hitler's desert function.

In this graph, points along the x-axis represent different possible levels of wellbeing at which Hitler could exist, and points along the y-axis represent different possible desert-values. For illustration, three points on the graph are labeled. Point α corresponds to a negative wellbeing level, call it level α. This is the peak of Hitler's desert function. If Hitler exists at level α, he gets what he perfectly deserves, and his existence has positive desert value [2] (chapter 4). All points to the left of point α correspond to wellbeing levels at which Hitler would overpay for his crimes. If Hitler were to exist at any of these levels, the desert value of his existence would be less than if he were to exist at level α. Point γ represents a positive wellbeing level for Hitler that corresponds to a negative desert value. If Hitler exists at level γ, then he gets the extreme opposite of what he perfectly deserves—he drastically underpays for his crimes. Point β represents a wellbeing level at which Hitler's existence has neutral desert value.

The existence of a neutral desert level seems to be a requirement of there being both some wellbeing level for Hitler with positive desert value and some wellbeing level for Hitler with negative desert value. It is hard to see how one could make sense of the distinction between positive and negative value without positing a neutral level that separates them. Moreover, it seems that an outcome in which Hitler exists at wellbeing level α, and that has positive desert value, and an outcome in which Hitler exists at wellbeing level γ, and that has negative desert value, can be connected along a spectrum of outcomes of descending desert values. For each outcome in this spectrum, the subsequent outcome differs from it only in that a small amount of the suffering that Hitler deserves is replaced with a small amount of pleasure, or something else that is intrinsically good for Hitler. As we descend from the first, good, outcome to the final, bad, outcome, things get better and better for Hitler, but worse and worse with respect to desert. The conclusion that some outcome in this spectrum is neutral with respect to desert seems inescapable. However, for my purposes, nothing of importance depends on *where* Hitler's desert function crosses the neutral level. The desert graph in Figure 1 shows Hitler's desert function crossing the neutral level at a point on the x-axis corresponding to a negative wellbeing level. However, readers can, if they like, imagine a different desert function for Hitler that is like the one in Figure 1 except that it crosses the neutral level right at the intersection of the x- and y-axes, or at some point to right of that intersection. What is important is that Hitler's desert function crosses the neutral desert level *somewhere*.

Are the wellbeing levels that correspond to different desert values *lifetime* wellbeing levels or *momentary* wellbeing levels? I will generally assume that they are lifetime wellbeing levels. This assumption will make it easier to see the parallels between the neutrality paradox (Section 3) and the desert-neutrality paradox (Section 4). Since the former paradox is presented in terms of lifetime wellbeing levels, it makes sense to present the latter in the same terms. However, the assumption that what a person deserves is a certain lifetime wellbeing is somewhat problematic. A retributivist could say that it is possible that an evil

person gets the suffering he deserves and yet has positive lifetime wellbeing. For example, suppose that a person becomes evil at a certain time in his life, commits certain wrongs, and is then punished at the very end of his life. Some retributivists might claim that if the punishment is fitting, then the suffering inflicted on the person is what he perfectly deserves. Yet, if the earlier parts of this person's life were sufficiently good for him, then his life might be good for him overall. In that case, although he would get the suffering that he perfectly deserves, he would also have positive lifetime wellbeing. This suggests that what he perfectly deserves is something more specific than a certain lifetime wellbeing, such as negative wellbeing during a certain period in his life. I address this potential problem for the assumption that what a person deserves is a certain lifetime wellbeing in Appendix A. There, I show that the desert-neutrality paradox arises even if one assumes that what evil agents deserve is not a life that is bad for them but rather a certain period of life that is bad for them.

In his introductory explanation of desert graphs, Kagan does not explicitly say whether such graphs can represent evaluative comparisons of a deserving person's existence and her non-existence [2] (chapter 3). However, he apparently thinks that desert graphs can represent such comparisons, since he applies the graphs to certain cases involving comparisons of outcomes in which different numbers of people exist [2] (pp. 601–619). In these cases, certain deserving people exist in some outcomes but not in others. Hence, comparisons of these different outcomes involve an implicit assumption about how a deserving person's existence compares to her non-existence. As I demonstrate in Section 4, such comparisons are problematic for axiological retributivism. Some such comparisons give rise to the desert-neutrality paradox, which parallels the more familiar neutrality paradox in population axiology.

3. The Neutrality Paradox

The neutrality paradox was introduced by John Broome and involves comparisons of outcomes in terms of their respective distributions of wellbeing, where some people who exist in certain of these outcomes do not exist in others [8] (pp. 143–185). In his presentation and analysis of the paradox, Broome does not consider how desert might affect the intrinsic value of outcomes. I will therefore bracket such considerations until Section 4, where I introduce the structurally similar desert-neutrality paradox and explore its implications for axiological retributivism. For the purposes of this section, I will assume that the values of the different outcomes that I will consider either do not depend on facts about desert or that these outcomes are all equally good with respect to desert and that therefore considerations of desert do not influence the comparison.

The neutrality paradox arises when one tries to develop a population axiology that can accommodate an axiological interpretation of what Broome calls the intuition of neutrality. The basic intuition is that there is a range of positive wellbeing levels such that the addition of a new person to the world at any of these levels is ethically neutral, apart from how it might impact the wellbeing of other people [8] (pp. 144–145). Broome provides examples that he considers evidence for the claim that some people might have this intuition. For instance, he points out that in cost-benefit analyses of projects to improve road safety, economists consider the expected number of lives saved but not the extra people who would exist since certain lives were saved. For instance, they ignore the fact that if a young person's life is saved, she will probably have children and grandchildren who would otherwise not have existed. One possible explanation of this is that the economists consider the addition of these extra people to be ethically neutral additions.

Broome acknowledges that there are different ways of interpreting the intuition of neutrality. One interpretation is deontic (i.e., related to moral obligation, duty, or reason). The economists estimating the benefits of improved road safety might think that their government has a moral obligation to protect its citizens, but no obligation to expand the population of its citizens by (indirectly) influencing the creation of new people, even if these new people would have good lives. Similarly, as Broome rightly points out,

couples considering whether to have a child are not morally obligated to do so, even if the child would have a good life, and even if the couple would be just as well off with or without a child [8] (p. 144). Some philosophers seem committed to a deontic interpretation of the intuition of neutrality [10,11]. For example, Jan Narveson argues that we have non-instrumental moral reason to improve the lives of those who exist, but no non-instrumental moral reason to create people with positive wellbeing. He expresses this idea with the slogan "We are in favor of making people happy, but neutral about making happy people" [10] (p. 80).

However, Broome is concerned mainly with an axiological interpretation of the intuition of neutrality. On this interpretation, the intuition is, roughly, that there is a range of positive wellbeing levels such that the existence of a person at any one of these levels does not make the outcome in which she exists better or worse than an outcome in which she does not exist, at least when we ignore the ways in which the extra person might impact the wellbeing of other people. A possible reason to think that the outcome in which the additional person exists is not better relates to the claim, which some find attractive, that an outcome cannot be better or worse (at least with respect to wellbeing) unless it is better or worse *for someone* [3]. Yet, it may seem that an outcome in which someone exists with a good life cannot be better for her than an outcome in which she does not exist. For this would seem to imply that the outcome in which she does not exist is worse for her [12,13]. But nothing can be worse (or better) for a person who does not exist. Putting all of these claims together, one would conclude that adding a well-off person to the world does not *by itself* make the world better.

The axiological interpretation of the intuition of neutrality expresses a claim about how additional people affect the value (goodness) of outcomes. It is a claim that is supposed to be part of a general theory of the value of outcomes with respect to their different distributions of wellbeing [8] (p. 145). However, depending on one's theory of the relationship between the right and the good, the axiological interpretation of the intuition could provide justification for the deontic interpretation. Indeed, one way of trying to justify Narveson's claim that we are neutral about making happy people is to appeal to the axiological interpretation of the intuition of neutrality. We might be neutral about making happy people since the addition of such people does not make the world better, other things being equal.

Broome argues that while the axiological interpretation of the intuition of neutrality may be attractive, one cannot accept it without rejecting at least one of several seemingly obvious axiological claims. This is the neutrality paradox. The paradox rests on some background assumptions about betterness [8] (pp. 21–22). It presupposes that betterness is a strict partial ordering. This means that the relation 'better than' is transitive (for all x, y, z, if x is better than y, and y is better than z, then x is better than z), asymmetric (for all x, y, if x is better than y, then y is not better than x), and irreflexive (for all x, it is not the case that x is better than itself). It also presupposes certain definitions. For instance, 'x is worse than y' means that y is better than x, and 'x is equally as good as y' means that x is neither better nor worse than y and any z is better (worse) than x iff z is better (worse) than y. It assumes that the relation 'equally as good as' is transitive, reflexive, and symmetric. All of these assumptions are prima facie plausible. But one could respond to the neutrality paradox by rejecting one or more of them. Temkin, for example, has explored doubts about the transitivity of 'better than' and 'equally as good as' [14] (chapters 6–7). However, I will not consider these doubts here, as they raise complicated issues that cannot be fully explored in this paper.

To see why it is difficult to accommodate the axiological intuition of neutrality, consider first the most natural interpretation of that intuition:

> *The principle of equal existence*. Suppose two outcomes have the same population, except that an extra person exists in one who does not exist in the other. Suppose each person who exists in both outcomes is equally as well off in one as she is in the other. Then there is some range of positive wellbeing levels ('the neutral

range') such that, if the extra person's wellbeing level is within this range, the two outcomes are equally good [8] (p. 146).

Broome presents what seem like decisive counterexamples to this principle. I will consider only Broome's first counterexample since it is the simplest. First, a clarificatory note. When considering Broome's example, as well as the examples that I shall present in Section 4, readers should not assume that the different possible outcomes in these examples are different alternatives, or options, for an agent. I will be considering how different outcomes compare in terms of their intrinsic value, not which outcome one would be obligated or permitted to bring about when facing a choice between different outcomes. If we imagine ourselves in a position to choose between the different outcomes that we are trying to evaluate, our intuitions may become distorted. We should focus only on the outcomes themselves and ignore the question of what someone ought to do when facing a choice between them.

Broome's simplest counterexample to the principle of equal existence involves a comparison of the three outcomes represented in Figure 2 [8] (pp. 146–148). The boxes in Figure 2 represent people. The height of a box represents the wellbeing level of the person it represents. The taller the box, the higher the wellbeing level, i.e., the better off the person is.

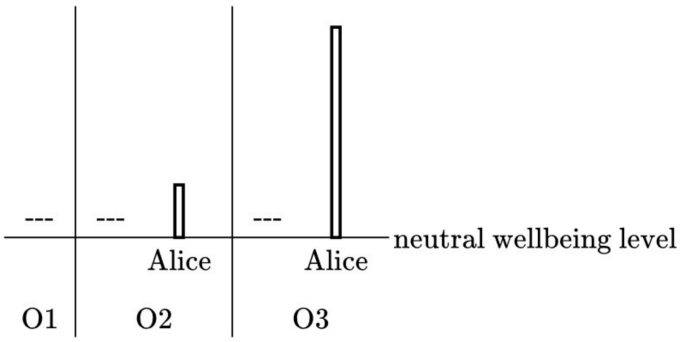

Figure 2. Counterexample to the principle of equal existence.

Alice does not exist in O1, but does exist in O2 and O3, and she has less wellbeing in O2 than in O3. The dashes in Figure 2 indicate that there are other people who exist in these outcomes besides Alice, but that one need not consider them for the purpose of the comparison. One can assume that their identities and wellbeing levels are fixed across O1, O2, and O3. The principle of equal existence states that there is a neutral range of wellbeing levels. Let us stipulate that Alice's wellbeing level in both O2 and O3 is within this range. According to the principle of equal existence,

(1) O2 is equally as good as O1 and
(2) O3 is equally as good as O1. From (1) and (2) and the transitivity of 'equally as good as',
(3) O2 is equally as good as O3. From (3) and the definition of 'equally as good as',
(4) O3 is not better than O2. But since Alice is better off in O3 than in O2, it seems obvious that
(5) O3 is better than O2.

Claims (1)–(5) are jointly inconsistent. Which should we reject? It seems clear that we should reject the two claims entailed by the principle of equal existence, (1) and (2), and thus block the inference to the ridiculous (3).

I have not presented the neutrality paradox in its entirety. As Broome points out, the neutrality of a person's existence can be interpreted in terms other than of equal goodness. It can be understood as *incommensurateness*. Two comparable things are incommensurate when one is neither better, worse, nor equally as good as the other [8] (pp. 165–171). It can also be understood in terms of the goodness of a person's wellbeing being conditional on her existence [8] (pp. 152–157). However, Broome exposes serious problems with these

alternative interpretations. He argues, for example, that the first interpretation is ad hoc and inconsistent with our intuitive idea of a person's existence being neutral with respect to goodness, and he argues that the second interpretation fails since the conditional goodness of a person's wellbeing cannot be fit into a coherent ranking of outcomes with respect to overall goodness. Thus, Broome ultimately rejects the axiological interpretation of the intuition of neutrality. In this section, I will not go into the details of Broome's discussion of the problems with the alternative interpretations of the neutrality of a person's existence. However, in Section 4, I will demonstrate that these very problems arise for axiological retributivism in the context of the desert-neutrality paradox.

Before proceeding to the discussion of the desert-neutrality paradox, it may be helpful to briefly reflect on what makes the neutrality paradox seem puzzling. Technically, the neutrality paradox arises due to the assumption that there are at least two wellbeing levels in the neutral range, and one is higher than the other. (It seems impossible to generate the paradox without this assumption.) However, in my view, what makes the neutrality paradox truly puzzling is something more fundamental, which can be expressed as follows. Some lives with positive wellbeing seem also to have positive value. In other words, some such life is not only good for the person who lives it but is also good absolutely. Now in some cases, such as Broome's case of cost-benefit analyses of road safety projects, such lives may seem like axiologically neutral additions. But how can the addition of a life with positive value, holding all other things fixed, be a neutral addition? How can the addition of a good life fail to make the outcome better when all other things are equal? There seems to be no satisfactory answer to this question. One possible answer is that there is some special kind of value interaction between the additional life and the other value-bearers that exist in the same outcome. However, it seems that we can just stipulate that in the cases of interest, there is no such special relationship between the additional good life and the other parts of the outcome that would explain how the addition of this good life could be neutral. But what else could explain how the addition of a good life fails to make the outcome better? That, I think, is the real puzzle for those who wish to accommodate the intuition of neutrality, and as we will see, the same kind of puzzle arises for axiological retributivism when this view is extended to comparisons of outcomes with different numbers of deserving agents.

4. The Desert-Neutrality Paradox

In this section, I show that axiological retributivism faces the desert-neutrality paradox. In terms of structure, this paradox is like the neutrality paradox, and the possible responses to the desert-neutrality paradox for retributivists parallel the possible responses to the neutrality paradox. However, there are two differences between these paradoxes. First, whereas the neutrality paradox concerns the goodness of an outcome with respect to its wellbeing distribution, the desert-neutrality paradox concerns the goodness of an outcome with respect to desert. I will, however, continue to assume that the betterness relation has the formal properties Broome claims it has, as well as Broome's definitions of 'equally as good as' and 'worse than'. In what follows, I will mostly dispense with the cumbersome locution 'better than with respect to desert' and will instead simply use 'better than'. However, readers should interpret instances of the latter as having the meaning of the former. The same goes for instances of 'worse than' and 'equally as good as'. Thus, readers should keep in mind that my claims in this section are about desert value, not all-things-considered value.

The second difference between the neutrality and desert-neutrality paradoxes is that the latter poses a bigger problem for retributivism than the former poses for population axiology. The neutrality paradox is not so paradoxical that it should lead to skepticism about population axiology or to the conclusion that there can be no adequate general population axiology. But I assume that the theoretical cost of abandoning population axiology altogether is much greater than that of simply rejecting the intuition of neutrality. However, I think, the desert-neutrality paradox does support skepticism about

axiological retributivism. Indeed, it seems to put significant pressure on axiologists to reject retributivism and retreat to a weaker position concerning the axiological significance of desert.

4.1. Retributivism and Comparisons of Existence and Non-Existence: Setting up the Paradox

In Section 1, we took Hitler as our example of an evil person who deserves what is bad for him (suffering), and we assumed a hypothetical desert function for Hitler, illustrated in Figure 1. The peak of the desert function is the negative wellbeing level α. This is what Hitler perfectly deserves. Although Hitler's existence at level α is bad for him it is good. Also represented in the desert graph is something that is good for Hitler but bad, namely existence at the very high positive wellbeing level γ, as well as something that is bad for Hitler but neutral, namely existence at the negative wellbeing level β, which is in between level α and level γ.

Our question is whether the desert graph in Figure 1 can plausibly represent evaluative comparisons of outcomes in which Hitler exists and those in which he does not exist. Let us first assign neutral desert value to a person's non-existence; that is, assume that the non-existence of a person by itself is neither good nor bad but neutral. (For now, we will refrain from any specific interpretation of the relevant sense of neutrality.)

Next, let us consider some comparisons. Suppose we compare outcome A, in which Hitler never exists, with outcome A-, in which Hitler exists at the very high positive wellbeing level γ. Suppose that apart from Hitler's existence, A and A- are equally good. This might be difficult to imagine. An outcome such as A, in which Hitler does not exist, would not have any of Hitler's evil deeds. That would seem to make A much better than A- in many respects, including desert. However, we can imagine that A has other evil deeds that do not occur in A-, and that the evil deeds in A and A- are equally bad, so that the only relevant difference between A and A- with respect to desert value is Hitler's existence in A- at level γ, and his non-existence in A. It seems Hitler's existence at level γ in A- is a bad-making feature of A- and that, since A and A- are otherwise equally good, A- is worse than A. More generally, if we are axiological retributivists, we should accept

> *The badness of additional undeserved good lives*. The addition of an evil person with a level of wellbeing that is much higher than what he perfectly deserves makes an outcome worse, other things being equal.

If we accept this claim, we should think that desert graphs as in Figure 1 plausibly represent at least some comparisons of a deserving agent's existence with his non-existence.

Next, compare A with a different outcome, A+, in which Hitler exists at the negative wellbeing level β. This is the level at which Hitler's existence has neutral desert value; he gets a lot of the suffering he deserves, but less than what he perfectly deserves. According to the graph, this is neither good nor bad but neutral. Again, we assume that apart from Hitler's non-existence in A and his existence at level β in A+, A and A+ are equally good. We should conclude that Hitler's existence at level β in A+ makes A+ neither better nor worse than A. With respect to desert, it is a neutral addition. This conclusion makes sense given our assumptions that Hitler's non-existence has neutral desert value and that his existence at level β also has neutral desert value.

Finally, compare A with yet another outcome, A++, in which Hitler exists at the negative wellbeing level α. An existence at level α is what Hitler perfectly deserves. Apart from Hitler's existence in A++ and his non-existence in A, these outcomes are equally good. How do A and A++ compare with respect to desert? We stipulated that Hitler's non-existence is neutral, and the desert graph in Figure 1 tells us that it is good if Hitler gets what he perfectly deserves. From these claims, we should conclude that A++ is better than A. An outcome in which Hitler exists and experiences all of the suffering he perfectly deserves is better than an outcome in which Hitler never exists, other things being equal.

But this conclusion is very counterintuitive. When I wear my retributivist hat, I find that A++ is not better than A.[3] If we are retributivists, we might think that if Hitler exists and commits certain evil deeds, then it is good if he experiences the suffering he perfectly

deserves. But is it really better in any respect if Hitler comes into existence and experiences that suffering? Would it really be worse if Hitler simply never existed?

In answering this question, we should not focus on the intrinsic badness of Hitler's atrocities. Our question is about the value of Hitler getting what he deserves. To focus your intuitions, consider a world in which Hitler gets all of the suffering he perfectly deserves. Suppose that in this world we also have a machine that can perfectly replicate Hitler as he was toward the end of his life. The replica would be just as evil as the original and would wholeheartedly endorse the original's wrongdoing. It would also be true of the replica that if he were, or had been, in Hitler's position, he would have committed all of the same atrocities. After creating the replica, we could ensure that it gets all of the suffering it perfectly deserves. Would it make things better in any respect to create the replica and punish it? I think not. This is not the case just due to the fact that the replica would (technically) be innocent of Hitler's actual crimes. Perhaps being extremely evil and wholeheartedly endorsing Hitler's crimes by itself warrants some form of punishment, such as (at least) a hard slap on the wrist. Still, I would not think that creating and punishing the replica would be, in any respect, an improvement. A retributivism that implies that it is better in some respect that an evil suffering person exists seems unpalatably sadistic. It goes too far.

But retributivists must also say that the addition of an evil person who gets the suffering he perfectly deserves does not make the outcome worse. For there is no retributivist rationale for the claim that the addition of such a person diminishes overall desert value. Hence, if we are axiological retributivists, we should want to accept

> *The neutrality of additional perfectly deserved bad lives*. The addition of an evil person with the negative wellbeing that he perfectly deserves makes an outcome neither better nor worse, other things being equal.

Unfortunately, retributivists cannot easily accommodate this claim. This is because they believe that some perfectly deserved lives with negative wellbeing have positive desert value. Accepting the claim stated above therefore puts retributivists in the difficult position of explaining how lives with positive desert value can be desert-neutral additions (i.e., additions that make the outcome neither better nor worse with respect to desert). This is exactly like the problem at the heart of the neutrality paradox considered in Section 2.

4.2. Introducing the Desert-Neutrality Paradox

Figure 3 represents the three outcomes described above, A, A+, and A++:

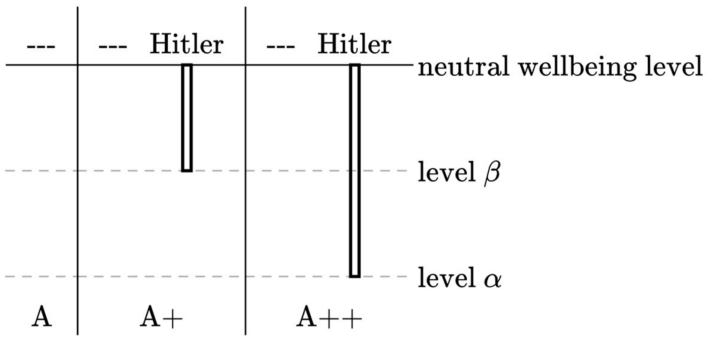

Figure 3. The desert-neutrality paradox.

In A Hitler doesn't exist, in A+ he exists at level β, and in A++ he exists at level α. In comparison with A, Hitler's existence in A+ is a desert-neutral addition. Given the neutrality of additional perfectly deserved bad lives, Hitler's existence in A++ is also a desert-neutral addition. These two claims jointly imply the existence of a range of wellbeing levels (call it 'the desert-neutral range') such that the addition of a specific person with a specific desert profile at either of these levels makes the outcome in which he exists neither better nor worse than an outcome in which he does not exist, other things being equal.

(The range has at least two wellbeing levels, level α and level β. However, it is plausible to assume that if levels α and β are included in the desert-neutral range, then any wellbeing level in-between levels α and β is also in that range.)

Suppose we interpret desert-neutrality in terms of equal goodness. Then

(6) A+ is equally as good as A and
(7) A++ is equally as good as A. By (6) and (7), and the transitivity of 'equally as good as',
(8) A+ is equally as good as A++. If (8) is true, then by definition,
(9) A++ is not better than A+. But axiological retributivism clearly entails
(10) A++ is better than A+.

Claims (6)—(10) are jointly inconsistent. Which should we reject? The answer seems to be (6) and (7), each of which follows from our interpretation of desert-neutrality as equal goodness.

But there are alternative interpretations of desert-neutrality—one in terms of incommensurateness and one in terms of conditional goodness. But for reasons of the kind that Broome identifies in the context of his discussion of the neutrality paradox, these alternative interpretations face serious problems. In the remainder of Section 4, we explore these problems.

4.3. Desert Neutrality as Incommensurateness

Suppose one rejects desert-neutrality as equal goodness. One can still claim that the addition of an evil person who gets what he perfectly deserves is neutral—that it makes things neither better nor worse. Ruth Chang, for example, has argued for the existence of a value relation that cannot be identified as one of the three standard relations, 'better than', 'worse than', and 'equally as good as' [15]. Retributivists might interpret desert-neutrality in terms of such a fourth value relation.

There are many alleged examples of a thing being neither better than, worse than, nor equally as good as another with respect to some evaluative dimension. Who was a better composer, Mozart or Debussy? [4] Some music lovers might favor one over the other. Yet, a reasonable person might, after careful reflection, decide that neither is better than the other, but also that the two are not *equally* good. When one considers the specific music-related values that make each composer great, for example Mozart's technical precision and Debussy's creative rebelliousness, one might be unable to make precise tradeoffs between these values, even if one were to somehow know all of the relevant facts about them. Following Broome, let us define 'x is incommensurate with y' (with respect to some value) to mean that x is neither better nor worse than y, nor equally as good as y (with respect to that value) [8] (p. 165).[5] Mozart and Debussy are incommensurate (with respect to their values as composers) just in case Mozart is neither better, nor worse, nor equally as good (a composer) as Debussy.

Two outcomes might be incommensurate with respect to desert. Is it better with respect to desert that a hundred evil people experience all of the suffering that they perfectly deserve or that one good person experiences most, but not quite all, of the joy and life satisfaction that she deserves? The comparison might be difficult to determine since the two outcomes differ both with respect to the number of deserving agents and with respect to the amounts of wellbeing that they deserve. Even if one were to know all of the relevant facts about the people in these two possible outcomes, the nature of their deeds, their specific vices and virtues, and so forth, one might find that neither outcome is better than the other with respect to desert and yet that the two outcomes are not equally good with respect to desert [2] (pp. 600–609). One might think that in this case there are two different desert values, the value of evil people getting what they deserve and the value of good people getting what they deserve, and that these values cannot always be precisely compared.

There is an important feature of incommensurateness that is crucial to the success of any solution to the desert-neutrality paradox that interprets desert-neutrality in terms of incommensurateness. This is that, unlike equal goodness, incommensurateness is not

transitive. That is, it is not the case that if x and y are incommensurate and y and z are incommensurate, then x and z are also incommensurate. (We will see why this matters shortly.) Notice that whenever it seems that one thing x is incommensurate with another thing y, it also seems that x is incommensurate with some third thing y+ that is clearly better than y. Suppose we think that a certain outcome O1, in which a hundred evil people experience all of the suffering they perfectly deserve, is incommensurate with an outcome O2, in which a single person gets most, but not all, of the good things that she perfectly deserves. Then we will probably also think that O1 is incommensurate with another outcome, O3, that is exactly like O2 except that the single good person in O3 has all of what she perfectly deserves in virtue of having slightly more of the good things she deserves than she has in O2. Clearly, O3 is better than O2. Yet O3 may still seem incommensurate with O1. A slight improvement in one of two incommensurate outcomes does not necessarily render the former better than the latter. But now we seem to have the following result: O2 and O1 are incommensurate, and O1 and O3 are incommensurate, but O3 is better than, and hence, not incommensurate with, O2.

We are now in a position to see how an appeal to incommensurateness might help retributivists avoid the desert-neutrality paradox. Suppose retributivists interpret the desert-neutrality claim in terms of incommensurateness. That is, suppose they say that the addition of a deserving person at a wellbeing level within the desert-neutral range results in an outcome that is incommensurate with an outcome in which this person does not exist, other things being equal. They can now avoid the problem described in Section 4.2. Recall that if we interpret desert-neutrality in terms of equal goodness, we must accept (6) A+ is equally as good as A, and (7) A++ is equally as good as A. These two claims, together with the transitivity of 'equally as good as' entail (8) A+ is equally as good as A++, which implies (9) A++ is not better than A+, which is inconsistent with retributivism. But if we interpret desert-neutrality in terms of incommensurateness, then we will reject (6) and (7) and replace them with

(6′) A+ and A are incommensurate and
(7′) A++ and A are incommensurate.

The conjunction of (6′) and (7′) does not imply (9). One cannot infer from (6′) and (7′) that A++ and A+ are incommensurate, since, as we just saw, unlike 'equally as good as', 'is incommensurate with' is not transitive.

However, this is not the end of the story. As Broome has shown, interpreting the neutrality of a person's existence as incommensurateness has problems. Broome illustrates these problems in the context of comparing different outcomes in terms of their different distributions of wellbeing [8] (pp. 169–170). But the same problems arise in the context of comparing different outcomes in terms of their desert values, as I shall now demonstrate.

One problem is that the appeal to incommensurateness is ad hoc. All of the examples that support the existence of incommensurateness involve comparisons in which one thing is better than another in some respect R1 but worse in some different respect R2. (Think of the legal and academic careers considered above.) It is because the tradeoff between R1 and R2 cannot be made with precision that we cannot say that either thing is better or that they are equally good. But comparisons of existence and non-existence are not like this. There is no reason to think that the addition of Hitler by itself leads to incommensurateness, apart from the fact that if this were true, we would avoid the problem in Section 4.2.

The second problem that Broome raises for treating the neutrality of personal existence as incommensurateness is that it fails to adequately capture our intuitive idea of neutrality:

> Suppose two things happen together. One is bad, and the other neutral. Intuitively, the net effect of the two things should be bad. A bad thing combined with a neutral thing should be bad. Intuitively, neutrality cannot act against badness to cancel it out, so the net effect should not be neutral [8] (p. 169).

Assuming that Broome would say the same thing about the combination of a good thing and a neutral thing, the general idea to which he appeals here is what I will call

The spirit of neutrality: A bad thing plus a neutral thing is a bad thing. Similarly, a good thing plus a neutral thing is a good thing.

According to Broome, an interpretation of the neutrality of a person's existence in terms of incommensurateness violates the spirit of neutrality, and hence, must be rejected.

To illustrate this violation in the context of desert value, consider the four outcomes represented in Figure 4:

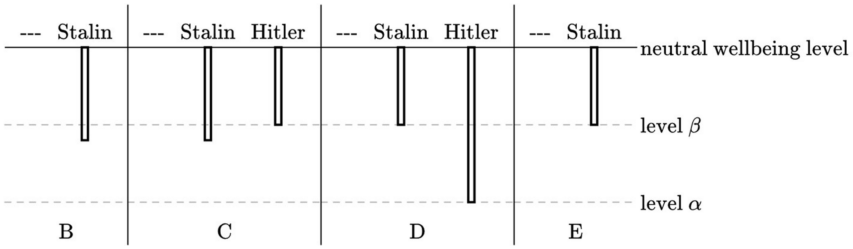

Figure 4. Desert-neutrality as incommensurateness.

This is similar to the previous example, except for the following difference. There is another evil character, Stalin, who exists in all four of the designated outcomes, B, C, D, and E. Stalin gets what he perfectly deserves in B and C; but in D and E, he exists at a wellbeing level, level β, that is slightly higher than what he perfectly deserves. Therefore, with respect to Stalin's desert, D and E are slightly worse than B and C. However, Hitler gets what he perfectly deserves in D, an existence at level α, so with respect to Hitler's desert, D is better than C. Let us stipulate that the welfare difference for Hitler between C and D has a greater impact on desert value than the welfare difference for Stalin, and that therefore with respect to desert,

(11) D is better than C and hence, by definition,
(12) C is worse than D. Retributivism clearly entails
(13) B is better than E.

(The only relevant difference between B and E is that Stalin gets what he perfectly deserves in B but not in E.)

The proposal we are considering is that desert-neutrality is incommensurateness. On this proposal, since wellbeing levels α and β are both within the desert-neutral range,

(14) C and B are incommensurate and
(15) D and E are incommensurate.

Here is where things get troublesome. Could D be worse than B? Apparently not. If D were worse than B, then from (12) and the transitivity of 'worse than' we could infer

(16) C is worse than B.

But then C and B couldn't be incommensurate as (14) states. Hence, we must instead accept

(17) D is not worse than B. Similarly, we must accept
(18) D is not better than B.

Why? Because the claim that D is better than B, as well as (13) and the transitivity of 'better than' jointly entail

(19) D is better than E.

But then D and E couldn't be incommensurate as (15) states. We cannot say that D is equally as good as B either, since from this claim and (13), and our definition of 'equally as good as', we get (19), which, as we just saw, contradicts (15). Hence, we must accept

(20) D is not equally as good as B.

Notice that (17), (18), and (20) together rule out the possibility of B and D standing in the relations 'better than', 'worse than', or 'equally as good as'. So, it seems we must accept

(21) B and D are incommensurate.

But (21) is inconsistent with the spirit of neutrality. D is worse than B with respect to Stalin's desert. We are assuming that the existence of Hitler in D makes D neither better nor worse than B. But these are the only two relevant differences between B and D. Hence, overall, D should be worse than B. A bad thing plus a neutral thing should amount to a bad thing. If D is incommensurate with B, despite being worse in one respect and neither better nor worse in another respect, then incommensurateness is, to use Broome's phrase, "greedy". It is a kind of neutrality that can "swallow up" bad things, or losses of good things.

One could respond to this objection to desert-neutrality as incommensurateness by rejecting the spirit of neutrality. For example, one might doubt Broome's claim that the spirit of neutrality captures our intuitive idea of neutrality. One reason to doubt Broome's claim is that our intuitive idea of neutrality seems logically compatible with the existence of organic unities, i.e., entities that possess a kind of holistic value that depends not only on the values of its parts but also the relations between these values. If there are organic unities, then presumably some are incompatible with the spirit of neutrality. Here is a putative example. If a certain person is in a Zen-like state and exhibits no emotional response whatsoever to what he observes, then this might be a neutral thing (neither good nor bad). If a certain child takes its very first steps, then this might be a good thing. But if a certain person exhibits no emotional response whatsoever to his child taking its first steps, then this may not be a good thing. It may instead be a bad thing, or perhaps a neutral thing. In this case, a certain neutral thing and a certain good thing might combine to yield something that is not good. The spirit of neutrality rules out the existence of such organic unities; it implies that a neutral thing and a bad (or good) thing cannot combine to yield anything other than a bad (or good) thing.

However, in the context of evaluating desert-neutrality as incommensurateness, this response is a red herring. Although Broome appeals to the spirit of neutrality, one need not appeal to this general claim to find fault with the interpretation of desert-neutrality as incommensurateness. One needs only the weaker claim that in the comparison of B and D, if the existence of Hitler in D is a neutral addition, and if D is worse than B with respect to Stalin's desert, and if other things are equal, then D is worse than B. This claim is only about two specific outcomes, B and D, so it does not rule out the existence of organic unities. Moreover, it is hard to see how the two differences just mentioned could fail to make D worse than B. Even if a bad (or good) thing and a neutral thing can sometimes amount to a neutral thing, we require an explanation of this wherever it happens. For example, in the case of the emotionless person observing his child taking its first steps, the explanation of how the neutral emotionless response and the good taking of first steps together amount to something neutral might be that there is some badness in failing to have the appropriate joyful response to the good event of one's child taking its first steps, and that this badness is equal in degree to the goodness of that event, such that the conjunctive state of affairs is neutral overall. But in the comparison of B and D, there is no explanation of this kind. Proponents of desert-neutrality as incommensurateness lack a principled explanation of desert-neutrality's greediness.

4.4. The Conditional Goodness of Deserved Suffering

Another option for retributivists is to try to explain the desert-neutrality claim in terms of the conditional goodness of deserved suffering. Conditional goodness is different from overall goodness. That x is good conditional on y entails that (y and x) is better than (y and not x) but (not y and x) is equally as good as (not y and not x), and moreover, (y and x) is equally as good as (not y and not x); in other words, the goodness of x is neutral with respect to whether the condition, y, obtains. As we saw earlier, this is how some philosophers, such as Narveson, think of a person's wellbeing [10]. They think that if a

person exists, it is better if she has more wellbeing; but they also think that it is not better if she exists in the first place and has positive wellbeing. The goodness of her wellbeing is conditional on her existence, and hence, is neutral with respect to whether she exists.

Retributivists can say something similar about the goodness of an evil person's suffering. They can say that this goodness is neutral with respect to whether the evil person exists. This avoids the uncomfortable result that it is better, other things being equal, if an evil suffering person comes into existence. Moreover, unlike the appeal to incommensurateness, the appeal to conditional goodness is not ad hoc.

However, there is a problem with understanding the goodness of deserved suffering as being conditional on an evil person's existence. It is basically the same problem that Broome raises for understanding the value of a person's wellbeing as conditional that person's existence [8] (pp. 152–157). It is hard to see how to fit the conditional goodness of deserved suffering into a coherent ordering of outcomes with respect to their overall desert value. To see this, consider the outcomes represented in Figure 5.

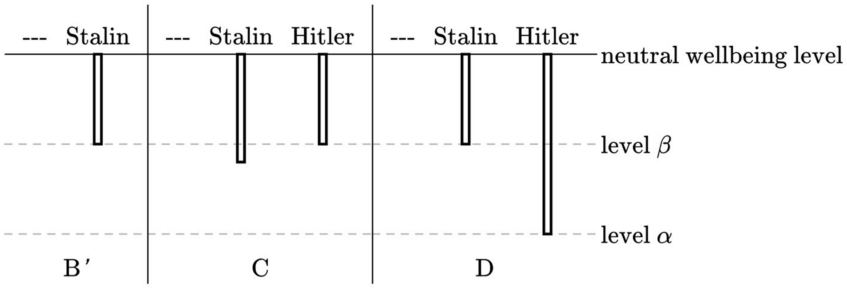

Figure 5. Conditional goodness of deserved suffering.

In Figure 5, as in Figures 3 and 4, level α and level β are denoted. Figure 5 also represents two of the outcomes featured in the previous case, C and D. B′ is similar to the outcome B featured in the previous case, except that in B′ Stalin is slightly better off than he is in B, hence, in B′ what Stalin gets is slightly better for him than what he perfectly deserves. As in the previous case, Stalin gets what he perfectly deserves in C and Hitler gets what he perfectly deserves in D. Moreover, as in the previous case, although D is worse than C with respect to Stalin's desert, it is better than C with respect to Hitler's desert, and we stipulate that the latter consideration outweighs the former, and hence,

(11) D is better than C. The conditional goodness of deserved suffering implies
(22) D is equally as good as B′.

The only difference between D and B′ is that Hitler exists in D and gets the suffering he perfectly deserves, but he does not exist in B′. The conditional goodness of Hitler's perfectly deserved suffering is neutral with respect to whether Hitler exists.

Since Hitler's deserved suffering is only a conditional good, with respect to Hitler's deserved suffering, C is equally as good as B′. But with respect to Stalin's deserved suffering, C is better than B′. It therefore seems that overall, with respect to desert

(23) C is better than B′. But (23) and (11), and the transitivity of 'better than' imply
(24) D is better than B′.

This contradicts (22), which follows from the conditional goodness of deserved suffering.

If 'better than' is transitive, then the pursuit of the conditional goodness of deserved suffering leads to a dead end. That is bad news for axiological retributivists, for it is plausible that if deserved suffering can be intrinsically good, then its intrinsic goodness is conditional on the existence of the deserving person.

5. Conclusions

I have argued that when axiological retributivism is extended beyond same-people cases it runs into the desert-neutrality paradox, a problem that mirrors the more familiar

neutrality paradox in population axiology. In this concluding section, I summarize the various ways out of the paradox for retributivists and offer some final thoughts about what the paradox might imply about moral decision-making.

One way out is to reject the neutrality of additional perfectly deserved bad lives, according to which the addition of an evil person who gets the suffering he perfectly deserves makes an outcome neither better nor worse with respect to desert, other things being equal. The problem here is that retributivists have no basis for claiming that the addition of such an evil person makes things worse with respect to desert, and the claim that such an addition makes things better with respect to desert leads to an unpalatably sadistic form of retributivism.

A second way out is to interpret a neutral addition of a deserving suffering evil person in terms of incommensurateness. The problem here is that the appeal to incommensurateness is completely ad hoc and seems incompatible with our natural understanding of axiological neutrality.

A third way out is to interpret a neutral addition of a deserving suffering evil person in terms of the conditional goodness of his deserved suffering. But retributivists will have to fit the conditional goodness of deserved suffering into a theory of overall goodness, and doing this leads to a violation of the transitivity of 'better than'.

A fourth way out, then, is to reject the transitivity of 'better than'. This proposal has generally been viewed as radical. However, recent work by Temkin suggests that most ethicists have not given this proposal a fair shake. Yet, even Temkin acknowledges that rejecting the transitivity of relations such as 'better than' and 'equally as good as' is hard to stomach, and could easily lead to skepticism about practical reasoning [14] (chapters 13–14).

A fifth proposal is not a way out but is a reasonable option. It is to reject axiological retributivism. If we reject retributivism, there are fallback positions we could accept, insofar as we want to recognize the axiological importance of desert. One would be the position considered in Section 1, according to which a certain evil person being closer (further), in terms of wellbeing, to what she perfectly deserves makes an outcome intrinsically better (worse) with respect to desert, but her getting the suffering (negative wellbeing) she perfectly deserves has intrinsic neutral value rather than intrinsic positive value, and her getting any more or less suffering than what she perfectly deserves is intrinsically bad. This fallback position seems to avoid the teeth of the desert-neutrality paradox since it implies that the only additions that are neutral with respect to desert are of evil agents who get what they perfectly deserve, and that any other additions of evil agents who deserve what is bad for them are bad with respect to desert. Those who accept this view must reject axiological retributivism's most central commitment, i.e., that deserved suffering has intrinsic goodness. But this may be a cost worth paying.

What, if anything, does the desert-neutrality paradox imply about morality? Although the desert-neutrality paradox is a paradox about the value of outcomes, many would acknowledge that the value of outcomes has implications for what agents ought morally to do, and for what they have moral reason to do. The specific implications of the desert-neutrality paradox for the moral status of acts and decisions will depend on one's theory of the relationship between the right and the good. It seems to me that on some of these theories, the desert-neutrality paradox will give rise to a corresponding moral paradox. This seems especially true of consequentialist moral theories on which the moral status of an act depends only on the values of the possible outcomes of the act.

Suppose that a consequentialist is committed to axiological retributivism. Then which of the different ways out of the desert-neutrality paradox should she embrace? The first way out, rejecting the neutrality of additional perfectly deserved bad lives, does not seem attractive. One who embraces this way out will have to say that creating an evil person who gets the suffering he perfectly deserves makes an outcome better with respect to desert, other things being equal. Thus, depending on how the consequentialist weighs the value of desert against other values, she might be committed to the claim that an agent would be

morally required to create such a person. For example, recall my example (Section 4.1) in which we have a machine that can create a replica of Hitler that would be just as evil as the original. Suppose that if we were to create the replica, we could allow it to reenact some of Hitler's evil deeds involving the infliction of suffering on innocent people. We would then punish the replica so that it experiences all of the suffering that it perfectly deserves. To offset the badness of innocent people's suffering we might promote the wellbeing of other innocent people, so that with respect to the general wellbeing of innocent people, the outcome in which the replica Hitler exists is neither better nor worse than the outcome in which it does not exist. Suppose that since these two outcomes would be equally good with respect to the general wellbeing of innocent people, the tie-breaking consideration would be the intrinsic goodness of the replica Hitler getting what he deserves. We might then conclude that the outcome in which the replica Hitler exists is better than the outcome in which it does not exist. If these outcomes are our best options, the consequentialist may embrace the conclusion that we are morally required to create and punish the replica, which is absurd.

Rejecting the transitivity of 'better than' is also an unattractive way out for consequentialists, especially if rejecting transitivity leads to value cycles, cases in which there are at least three options, x, y, and z, such that x is better than y, y better than z, and z better than x. Insofar as consequentialism implies that it is wrong to choose an option that is worse than some other option one could instead choose, value cycles would lead to moral dilemmas—cases in which an agent will act wrongly no matter what she does [17,18]. Value cycles would also make it impossible to maximize the good in every morally relevant situation, since this seems to require that there is always at least one option that is not worse than any other option.

Given these problems for the combination of consequentialism and axiological retributivism, consequentialists may have an additional reason to reject axiological retributivism, apart from any reason that has to do with the problematic nature of the desert-neutrality paradox itself. It is difficult to say whether non-consequentialist moral theories would face similar problems. It seems possible, albeit not inevitable, that any moral theory that recognizes the moral significance of the goodness of outcomes will, when combined with retributivism, encounter such problems.

Funding: This research was funded by Riksbankens Jubileumsfond, grant number M17–0372:1.

Acknowledgments: For helpful discussion of the desert-neutrality paradox, I am grateful to Krister Bykvist and Steve Kershnar.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A Momentary Wellbeing and the Desert-Neutrality Paradox

In Section 1, I claimed that the desert-neutrality paradox doesn't depend on the assumption that what evil people deserve is negative lifetime wellbeing level. This appendix demonstrates that claim.

The examples in this paper involve pairs of outcomes that differ in that one has an evil agent who does not exist in the other and who gets the negative wellbeing he perfectly deserves. A retributivist might think that what is deserved in these cases is not negative lifetime wellbeing but negative wellbeing at some temporal interval in one's life. So, instead of comparing outcomes that differ only with respect to the existence or non-existence of an evil agent with negative lifetime wellbeing, we can instead compare outcomes that differ only with respect to the existence or non-existence of an evil agent whose life contains the temporary negative wellbeing (suffering) that he perfectly deserves. To illustrate, consider Figure A1 below. Rather than represent different people who exist in different outcomes, Figure A1 represents different time-slices of a possible life of Hitler. In this case, each box represents a temporal interval in Hitler's life, and the height of the box represents how well Hitler fares during that interval—the higher a box extends above the neutral temporal wellbeing level, the better Hitler fares during the temporal interval represented by that

box, and below this level, the lower the box extends, the worse Hitler fares during the interval represented by that box. The width of a box represents the duration of the relevant interval—the wider the box, the longer the duration.

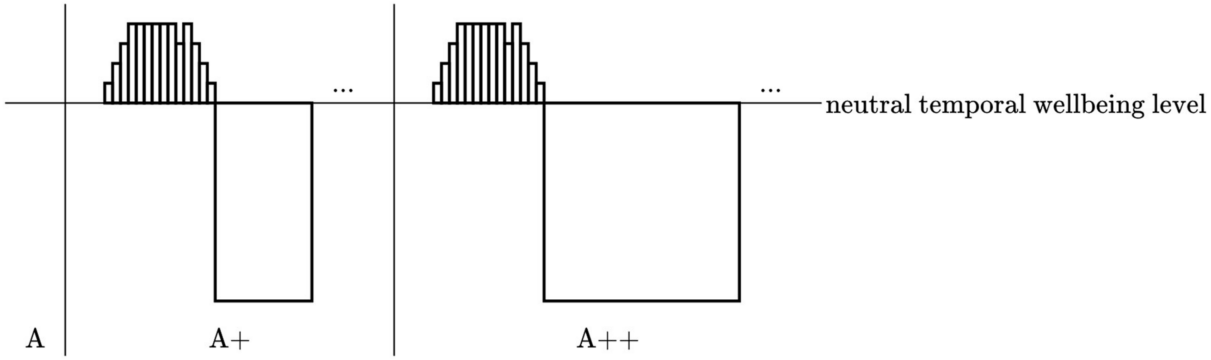

Figure A1. Time-slices of Hitler's life.

Three outcomes are represented, A, A+, and A++. In A, Hitler does not exist. In A+ Hitler exists and his life goes better and worse at different times, but during one long period he experiences a lot of the suffering he deserves, but less than the amount of suffering that he perfectly deserves. This period is represented by the large box in A+ that sits below the neutral temporal wellbeing level. We stipulate that A+ has neutral desert value. In A++, Hitler exists and gets the prolonged suffering that he perfectly deserves, which is more suffering than he gets in A+. The ellipses to the right of the large boxes in A+ and A++ indicate that Hitler's life continues after his prolonged suffering. We can assume that the part of his life after his deserved suffering is good for him and that his life is good for him overall. We stipulate that the parts of Hitler's life that are not represented by the boxes in Figure A1 do not make a relevant difference to how A, A+, and A++ compare. In other words, apart from the contribution that Hitler's deserved suffering makes to the value of these outcomes, other things are equal.

This version of the desert-neutrality paradox has the same structure as the version presented in Section 4.2. Hitler's non-existence is neutral with respect to desert. Hitler's existence in A+ is a neutral addition. Hitler's existence in A++ is better with respect to desert than his existence in A+, since in A++ he gets the amount of suffering that he perfectly deserves, which is better than his getting less than that amount. But is A++ really better than A with respect to desert? If it is, then Hitler coming into existence and experiencing the suffering he perfectly deserves makes an outcome better, other things being equal. This gives retributivism an unpalatably sadistic character.

Notes

1 Temkin presented this thought experiment in a seminar that I attended at Rutgers University in the fall of 2021. See also [3]

2 Kagan uses the term 'absolutely deserve' rather than my term 'perfectly deserve', but these terms are synonymous

3 Arrhenius expresses a similar intuition. See [8] (p. 234)

4 This example is inspired by one of Parfit's. See [16] (p. 113–114)

5 I have added the qualification 'with respect to some value', which is not part of Broome's definition

References

1. Zaibert, L. *Rethinking Punishment*; Cambridge University Press: Cambridge, UK, 2018.
2. Kagan, S. *The Geometry of Dessert*; Oxford University Press: New York, NY, USA, 2012.
3. Temkin, L. Harmful Goods, Harmless Bads. In *Value, Welfare, and Morality*; Frey, R.G., Morris, C.W., Eds.; Cambridge University Press: Cambridge, UK, 1993; pp. 290–324.
4. Feldman, F. Justice, Desert, and the Repugnant Conclusion. *Utilitas* **1995**, *7*, 189–206. [CrossRef]
5. Feldman, F. Adjusting utility for justice: A consequentialist reply to the objection from justice. In *Utilitarianism, Hedonism, and Desert: Essays in Moral Philosophy*; Cambridge University Press: Cambridge, UK, 1997; pp. 151–174.

6. Broome, J. *Weighing Lives*; Oxford University Press: Oxford, UK, 2004.
7. Rabinowicz, W. Broome and the Intuition of Neutrality. *Philos. Issues* **2009**, *19*, 389–411. [CrossRef]
8. Arrhenius, G. Feldman's Desert-Adjusted Utilitarianism and Population Ethics. *Utilitas* **2003**, *15*, 225–236. [CrossRef]
9. Arrhenius, G. Desert as Fit: An Axiomatic Analysis. In *The Good, the Right, Life and Death: Essays in Honor of Fred Feldman*, 1st ed.; Raibley, J.R., Zimmerman, M.J., McDaniel, K., Eds.; Routledge: New York, NY, USA, 2006; pp. 3–17.
10. Narveson, J. Moral Problems of Population. *Monist* **1973**, *57*, 62–86. [CrossRef] [PubMed]
11. Frick, J. Conditional Reasons and the Procreation Asymmetry. *Philos. Perspect.* **2020**, *34*, 53–97. [CrossRef]
12. Persson, I. *Inclusive Ethics: Extending Beneficence and Egalitarian Justice*; Oxford University Press: Oxford, UK, 2017.
13. Bykvist, K.; Campbell, T. Persson's Merely Possible Persons. *Utilitas* **2020**, *32*, 479–487. [CrossRef]
14. Temkin, L. *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*; Oxford University Press: New York, NY, USA, 2012.
15. Chang, R. The Possibility of Parity. *Ethics* **2002**, *112*, 659–688. [CrossRef]
16. Parfit, D. Can We Avoid the Repugnant Conclusion? *Theoria* **2016**, *82*, 110–127. [CrossRef]
17. Sinnott-Armstrong, W. *Moral Dilemmas*; Basil Blackwell: Oxford, UK, 1988.
18. Arrhenius, G. Can the Person Affecting Restriction Solve the Problems in Population Ethics? In *Harming Future Persons: Ethics, Genetics, and the Non-Identity Problem*; Roberts, M., Wasserman, D., Eds.; Springer: New York, NY, USA, 2009.