1. Introduction

Population ethics is often described as a field littered with paradoxes. In this paper, I sketch a novel solution to one of the most venerable and interesting puzzles in this branch of moral philosophy – Derek Parfit’s Mere Addition Paradox.1 The paradox is of enduring interest for several reasons. For one, it can be read as one of the first “impossibility results” in population ethics. It suggests that many people’s pre-theoretic convictions about how we ought to shape the size and quality of life of future generations are in deep tension with one another. Worse than that, a natural extension of the paradox appears to show that no theory of population ethics can avoid a result that strikes many as profoundly counterintuitive: the dreaded Repugnant Conclusion. Yet Parfit’s case is more than just an arcane puzzle, of interest to the field of population ethics alone. Over the past three decades, philosophers’ attempts to grapple with this problem have brought forth some of the most original and daring work in the theory of value, work which challenges some of the most basic assumptions about the nature of the good itself, in particular the notion that the “all-things-considered better than”-relation is transitive.2

In this paper, I aim to dissolve the Mere Addition Paradox. The paradox, I argue, trades on an ambiguity about the context of choice. Once this ambiguity is removed, we see that on one way of specifying the context of choice, there is in fact no tension between the three pairwise judgments that make up the paradox, whereas on another way of specifying the context of choice, we lose our basis for affirming one of its three constituent propositions. In neither case does the Mere Addition Case threaten to take us to the Repugnant Conclusion. The lessons for value theory in general are also less dramatic than some had thought. Correctly understood, Parfit’s

---

1 Though Parfit had discussed versions of the Mere Addition Paradox in earlier work, his canonical presentation of the problem, on which I draw in what follows, is in Chapter 18 of Reasons and Persons (Oxford: Oxford University Press, 1984).

Mere Addition Case challenges not the transitivity of the “better than”-relation, but instead a different, and less sacrosanct, idea, namely the so-called Independence of Irrelevant Alternatives principle.

2. The Mere Addition Paradox

Parfit presents the Mere Addition Paradox in the context of searching for what he called “Theory X”, that is a complete and satisfactory axiology for the field of population ethics. Such a population axiology, if one could be found, could then be fed into a principle of beneficence, which would tell us, alone or in conjunction with other moral principles, how to act in shaping the size and quality of life of future generations.

One very simple candidate for Theory X, which is embraced by classical utilitarianism, is what Parfit calls

**The Impersonal Total Principle**: “If other things are equal, the best outcome is the one in which there would be the greatest quantity of happiness—the greatest net sum of happiness minus misery.”

The Impersonal Total Principle implies that a loss in the quality of lives in a population can be compensated for by a sufficient gain in the quantity of lives that are lived. However, this consequence of the Impersonal Total Principle gives rise to a well-known problem. It implies

**The Repugnant Conclusion**: “For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better, even though its members have lives that are barely worth living.”

---

3 Following Hilary Greaves, I understand a population axiology as “a betterness ordering of states of affairs, where the states of affairs include ones in which different numbers of persons are ever born.” See Hilary Greaves, “Population Axiology”, *Philosophy Compass* 12(11) (2017), p. 1.


This is a conclusion which most philosophers working in the field of population ethics, Parfit included, find intuitively unacceptable. Someone who embraces the Impersonal Total Principle, however, can be forced to the Repugnant Conclusion via a simple “Spectrum Argument”, illustrated by Figure 1:

![Figure 1](image)

Each block represents a different possible history of the world, with the width of the block representing the number of people who ever live and the height of the block their quality of life. Someone who accepts the Impersonal Total Principle, it is argued, must grant that B is better than A. For, while the people in B are slightly less happy than the people in A, there are sufficiently more of them to make B all things considered better than A. Mutatis mutandis for the comparison between C and B. Iterated application of this step, plus the transitivity of “all-things-considered better than”, lead us to the Repugnant Conclusion that Z, an extremely large population made up of people whose lives are barely worth living, is better than A, a population of 10 billion people with very happy lives.

It might be thought that we can easily escape this predicament by rejecting the Impersonal Total Principle. Many philosophers, myself included, are attracted to an alternative view, which John Broome has dubbed

**The Intuition of Neutrality**: While it makes the world go worse, all else equal, to add to it a life that is foreseeably not worth living, it does not make the world go better, all else equal, to add to it a life that is worth living. Rather, doing so is ethically neutral.7

---

6 For a dissenting voice, see Torbjörn Tännsjö, ‘Why We Ought to Accept the Repugnant Conclusion’, *Utilitas* 14 (3) (2009), pp. 339–359.

The Intuition of Neutrality neatly dovetails with a deontic intuition about the morality of procreation, the so-called Procreation Asymmetry, which strikes many as powerfully attractive. According to this intuition, if a future person would foreseeably have a life that is not worth living, this in itself gives us a strong moral reason to refrain from bringing this person into existence. By contrast, there is no moral reason to create a person whose life would foreseeably be worth living, just because her life would be worth living.\(^8\) If, on the other hand, we deny the Intuition of Neutrality and embrace the view that each additional happy life makes the world better all else equal, the Procreation Asymmetry may seem hard to defend. If creating a new happy life would make the world better, all else equal, shouldn’t we have some moral reason to do so?

Rejecting the Intuition of Neutrality would have strongly revisionary implications in many other contexts as well. A nice example comes from John Broome. Considering the moral value of programs for improving people’s safety, he writes:

When people’s lives are saved, by making roads safer or in other ways, the wellbeing of the people who are saved is generally small in comparison to the wellbeing of all the new people, their descendants, who come into existence as a result. This is perfectly predictable. If all the descendants’ wellbeing had to be counted too, that would enormously alter the value we attach to saving people’s lives. But actually, in judging the value of safety on the roads, we routinely ignore all this wellbeing.\(^9\)

If we accept the Intuition of Neutrality, then the Spectrum Argument towards the Repugnant Conclusion never gets off the ground, since we reject its very first step. If containing more happy lives does not make the world go better all else equal, there is no reason to accept that B is better than A (nor that C is better than B, etc.). So the argument never gets going.

---


Parfit’s Mere Addition Paradox shows that, unfortunately, avoiding the Repugnant Conclusion is not as simple as that. Even if we reject the Impersonal Total Principle and embrace the Intuition of Neutrality, there is a second argument that threatens to take us to the Repugnant Conclusion all the same. Consider Figure 2:

![Figure 2](image)

Figure 2

A represents a population of 10 billion people, all with very happy lives. A+ consists of two groups. The first group consists of the same people as A, with the same high quality of life (let’s call these the “A-People”). A+ also contains a group of extra people, equal in number to the A-People. (Let’s call these the “New People”). These New People are significantly worse off than A-People, but still have lives that are well worth living. Outcome B contains the same two groups as A+. Everyone in outcome B is worse off than the people in A, but better off than the New People in A+. Assume that the average quality of life in A+ (and hence also the total wellbeing in A+) is below that in B.

In *Reasons and Persons*, Parfit very convincingly argues that each of the following three pairwise comparisons is extremely plausible:

1. A+ is worse than B.
2. B is worse than A.
3. A+ is not worse than A.

Let us briefly review the case for each of these three claims.

The intuition that A+ is a worse outcome than B is for many people the most robust. It is also supported by a number of distinct moral views: Total utilitarians, average utilitarians, telic egalitarians, and prioritarians all converge on the claim that B is a better outcome than A+.
The second claim, that B is worse than A, also strikes many people as strongly intuitive. *Everyone* who exists in B is worse off than everyone who exists in A (including the people who exist in both A and B). This verdict would, of course, be contradicted by proponents of the Total Principle. But as we have seen, the Total Principle itself is suspect, since it implies the Repugnant Conclusion. By contrast, for someone attracted to the Intuition of Neutrality, the claim that B is worse than A should be especially attractive. For a move from A to B can be decomposed into two distinct effects, as Figure 3 illustrates.

First, there is a change in the well-being of the A-People (highlighted in purple). This is the move from A to B₁. B₁ is incontrovertibly worse than A. Second, there is the addition of the New People, B₂, to B₁. However, if the Intuition of Neutrality is correct, adding new lives worth living does not make the world go better. So, B, the union of B₁ and B₂, is not better than B₁. And B₁, we said, is worse than A. So, all things considered, B is worse than A.

Finally, consider the comparison between A and A+, represented by Figure 4.
Relative to $A$, $A+$ involves what Parfit calls “mere addition”, meaning that the existence of the New People in $A+$ does not affect anyone else’s well-being, and that they have lives that are worth living.

If we reject the Impersonal Total Principle and accept the Intuition of Neutrality, we don’t claim that the existence of some additional people with lives worth living makes the outcome better, compared to $A$. But, at the same time, Parfit thinks it is very hard to see how merely adding a set of new people could make an outcome worse. We should embrace

**The Mere Addition Principle**: Merely adding a set of new people, all of whom have lives that are worth living, does not make an outcome worse.

This principle might be challenged: What about the fact that there is inequality in $A+$, whereas there isn’t in $A$? Couldn’t this fact be thought to make $A+$ worse than $A$? To this *Objection from Inequality*, Parfit gives the following response in *Reasons and Persons*: We should distinguish two kinds of case: If we can avoid inequality by making some existing people better off, then avoiding inequality can make the outcome better. But if inequality can only be avoided if some people, who have lives worth living, never exist, then avoiding inequality does not make the outcome better. So, Parfit concludes, when inequality is produced by mere addition, it does not make the outcome worse. As he puts it: “We cannot plausibly claim that the extra people should never have existed merely because […] there are other people who are even better off.” If this argument goes through, then the inequality in $A+$ gives us no reason to reject the thought that $A+$ is not worse than $A$.

But now it seems we have a problem: We have affirmed that $A$ is better than $B$, and that $B$ is better than $A+$. By the transitivity of “better-than”, it seems we are committed to the claim that $A$ is better than $A+$. But the Mere Addition Principle denies this.

We therefore have a set of three propositions, each of which intuitively compelling, but which apparently cannot all be true – a paradox.

By itself, this simple Mere Addition case is not enough to take us to the Repugnant Conclusion. For the least implausible revision to our three pairwise judgments, Parfit thinks, is to revise the claim that $A$ is better than $B$ to the claim that $B$ is not worse than $A$. We are then left with the following three pairwise judgments:

---

(1) A+ is worse than B.
(2) B is not worse than A.
(3) A is not worse than A+.

These three judgments are logically consistent, since the “not worse than”-relation, unlike the “better than”- or “worse than”-relations is not transitive.\textsuperscript{11} And there is no argument from the conclusion that B is \textit{not worse than A} to the Repugnant Conclusion that population Z is \textit{better than A}.

There is, however, a “supercharged” variant of the Mere Addition Paradox which does threaten to take us all the way to the Repugnant Conclusion. In what Parfit calls the “Up-Down Case”, illustrated by Figure 5, the A-People are actually made \textit{better-off} when we add the New People (hence the labeling of this outcome as “A++”). Of course, even in this case we are to assume that average utility in A++ is lower than in B, which in turn has a lower average utility than A.

\begin{figure}
\centering
\includegraphics[width=0.5\textwidth]{figure5.png}
\caption{Figure 5}
\end{figure}

As in the original Mere Addition Case, Parfit thinks it is very plausible that

(1) B is better than A++.
(2) A is better than B.

\textsuperscript{11} For example, we might think that a career as a lawyer and a career as a musician are qualitatively sufficiently different to vitiate very precise evaluative comparisons. It may then be the case that a career as a lawyer, earning $150k a year, is \textit{not worse} than a career as a musician, earning $80k a year, and that a career as a musician, earning $80k a year, is \textit{not worse} than a career as a lawyer earning $160k a year. But a career as a lawyer, earning $150k a year, is strictly \textit{worse} than a career as a lawyer, earning $160k a year.
The same considerations in support of these claims that we just surveyed still apply.

However, in the Up-Down Case, not only is A++ not worse than A. We should affirm a stronger claim:

(3) A++ is better than A.

After all, Parfit reasons, had we chosen A instead of A++, this would have been worse for everyone who exists in A; and in addition, a second group of people, all of whom have lives that are well worth living, would never have existed. Given this, it is hard to see how A would not have been worse than A++.

So, in the Up-Down Case, it seems very plausible that A++ is better than A, and that B is better than A++. But since “better than” is a transitive relation, these two plausible judgments entail that B is better than A. And this reasoning can now be iterated. By the same two step Up-Down procedure, we can argue that B++ is better than B, that C is better than B++, and so on – all the way to the Repugnant Conclusion.

These, in brief, are Parfit’s claims about the Mere Addition Paradox and the Up-Down Case. The significance of these problems for population axiology are self-evident: If Parfit’s arguments about the Mere Addition Case are sound, Theory X will not be able to accommodate all our pretheoretic convictions concerning population axiology, since these are inconsistent. Moreover, unless the challenge of the Up-Down Case can be met, we must abandon the hope of constructing a Theory X that avoids the Repugnant Conclusion.

2. Temkin on the Mere Addition Paradox and the Non-Transitivity of “Better Than”

The Mere Addition Paradox, however, is far more than just a theoretical puzzle for population ethicists. In his seminal paper “Intransitivity and the Mere Addition Paradox”12 and his book Rethinking the Good13, Larry Temkin forcefully argues that fully coming to grips with this problem should

---

prompt us to reexamine some of our most fundamental assumptions about the nature of the good itself.

Temkin argues that, besides accepting that at least one of the intuitively compelling judgments (1) through (3) is false, there is a fourth possible solution to the paradox that needs to be considered. This is to reject the idea that the “all-things-considered better-than”-relation is transitive.

To many, of course, this proposal will seem a non-starter. Surely, if there is one thing we know about the nature of the good, it is that “better than” is transitive.14

Temkin concedes that this idea will seem rationally undeniable, if we accept a certain picture of the nature of our moral ideals – what he calls the “Internal Aspects View of Outcome Goodness” (IAVOG). According to this view,

[r]oughly, for each outcome, O, how good that outcome is all things considered depends solely on how good it is with respect to each moral ideal that is relevant for assessing the goodness of outcomes, and on how much all of the relevant ideals matter vis-à-vis each other, where these depend solely on O’s internal features. Moreover, for any two outcomes, O₁ and O₂, O₁ will be better than O₂ all things considered if and only if the extent to which O₁ is good all things considered, as determined solely on the basis of O₁’s internal features, is greater than the extent to which O₂ is good all things considered, as determined solely on the basis of O₂’s internal features.15

If all our moral ideals which together determine the goodness of an outcome conform to the Intrinsic Aspects View, then we can assign each outcome a goodness “score”, which is solely a function of its “internal”, or intrinsic, properties. And then the goodness of outcomes surely must be transitive. If outcome 1 has a higher goodness score than outcome 2, and outcome 2 has a higher goodness score than outcome 3, and these scores are solely a function of the intrinsic features of these three outcomes, then the score of outcome 1 must also be higher than that of outcome 3.

However, Temkin argues that if we are truly compelled by Parfit’s reasoning in support of each of the three pairwise judgments that make up

the Mere Addition Paradox, then this should lead us to question the Internal Aspects View.\textsuperscript{16} What may underlie and justify our non-transitive intuitions in the Mere Addition Paradox, Temkin argues, is the fact that certain of our moral ideals, such as the ideal of equality, may be what he calls “essentially pairwise comparative”. For instance, Temkin maintains that on the view suggested by Parfit’s claims about the badness of inequality, it is true that “[equality is] comparative, not merely in the ordinary sense – that it involves judgments about how some fare relative to others – but in the sense that our judgment about a situation’s inequality depends on the alternative it is being compared to.”\textsuperscript{17} On this view about equality\textsuperscript{18}, Temkin writes, “the relevant and significant factors for comparing A and A+ regarding inequality differ from those for comparing A+ and B in a sense connected with inequality being essentially pairwise comparative.”\textsuperscript{19} Thus, “[t]here is no fact of the matter as to how bad the inequality in A+ really is considered just by itself. How bad it is depends upon the alternative compared to it. Compared to B, A+ is bad; compared to A, it isn’t.”\textsuperscript{20}

Hence, according to Temkin, if we accept Parfit’s reasoning about the Mere Addition Paradox, this suggests the following lesson: Rather than being a function only of an outcome’s intrinsic aspects, the overall goodness of an outcome is often essentially comparative or as I will say comparison-dependent. Depending on which alternative it is being compared to, an outcome will do better or worse in terms of certain moral ideal, such as equality. And since these ideals are part of what makes an outcome better or worse all-things-considered, the all-things-considered goodness of outcomes will often be comparison-dependent as well. If all this is true, we should not expect the “all-things-considered better than”-relation to be transitive.

\textsuperscript{16} Temkin does not claim that our intuitions about the Mere Addition Paradox require us to give up on the transitivity of “all-things-considered better than”. We may instead choose to give up one of our considered judgments about the Mere Addition Paradox and preserve the transitivity of the better than-relation. What Temkin maintains is that if we do embrace certain deeply plausible ways of thinking about the Mere Addition Paradox (amongst them Parfit’s claims in Reasons and Persons about when inequality makes an outcome worse and when it doesn’t, which we surveyed above, but also certain forms of Maximin, as well as the Narrow Person-Affecting view), then this would force us to accept that the “all-things-considered better than”-relation is indeed intransitive in a deep sense. See Rethinking the Good, Chapter 12.

\textsuperscript{17} Temkin, ‘Intransitivity and the Mere Addition Paradox’, p. 147.

\textsuperscript{18} Note that this is not Temkin’s own view. On Temkin’s own view, equality is not essentially pairwise comparative. Indeed, the view about equality that Temkin defends in Inequality (New York: Oxford University Press, 1993) is compatible with the IAVOG.

\textsuperscript{19} Temkin, ‘Intransitivity and the Mere Addition Paradox’, p. 149.

\textsuperscript{20} Temkin, ‘Intransitivity and the Mere Addition Paradox’, p. 150.
I believe that Temkin’s remarks on the Mere Addition Paradox contain, at once, a deep insight and an important error. I think Temkin is right that if we take Parfit’s claims about the Mere Addition Paradox seriously, they present a potent challenge to the Internal Aspects View of Outcome Goodness. However, we should depart from the Internal Aspects View in a different way than Temkin suggests. The lesson of the Mere Addition Paradox, I claim, is not that the goodness of outcomes is often comparison-dependent, but that it is context- or choice-set dependent. Accordingly, we shall see that what the Mere Addition Paradox challenges is not the transitivity of “better than”, but a different principle of rational choice, the Independence of Irrelevant Alternatives.

3. A Crucial Ambiguity

The Mere Addition Paradox, I claim, trades on an ambiguity about the context of choice. As others before me have remarked, the three pairwise comparisons which Parfit makes in building up the Mere Addition Paradox implicitly assume a particular context of choice, namely one in which the feasible set contains only the two options being compared. Let us call this a two-possible choice.

That Parfit must be making this assumption is clear from the way he deals with the Objection from Inequality against the claim that A+ is not worse than A (or indeed better than A, in the Up-Down Case). Parfit’s response to this objection, recall, was that if the only way of removing the inequality between the A-People and the new people is for the latter group never to exist, then the inequality is not a bad-making feature of A+. This is true in a two-possible choice between A and A+, and supports the judgment that, in that case, A+ is not worse than A. To cite one of Parfit’s own examples, we might imagine that the new people in A+ are all born blind, in a world in which there is no cure for blindness.

By contrast, in a 3-possible choice, where A, A+, and B are all in the feasible set, this argument would not get a purchase. For here there does exist a way of removing the inequality between the A-People and the New People other than by preventing the New People from coming into existence, namely option B. Imagine that the New People are born blind into a society which

---

does have a cure for blindness. If the seeing (the A-People) were willing to sacrifice some of their resources, the blind could all be cured of their disability.

I claim that there is a sense in which all three intuitive judgments about Parfit’s Mere Addition Case are true, namely as pairwise comparisons in two-possible cases. Thus, it seems to me true that

(1*) A is better than B, when the option set is \( Q = \{ A, B \} \);
(2*) B is better than \( A+ \), when the option set is \( R = \{ A+, B \} \);
(3*) \( A+ \) is not worse than A, when the option set is \( T = \{ A, A+ \} \).

The air of paradox arises from the assumption that, if these three pairwise rankings are true in two-possible choices, they all carry over to the three-possible case, in which the feasible set contains all three outcomes. It would then be true that

(1’) A is better than B, when the option set is \( S = \{ A, A+, B \} \);
(2’) B is better than \( A+ \), when the option set is \( S = \{ A, A+, B \} \);
(3’) \( A+ \) is not worse than A, when the option set is \( S = \{ A, A+, B \} \).

This would spell big trouble, since (1’), (2’), and (3’) would together constitute an instance of what I call “non-transitivity within an option set” as opposed to the “non-transitivity across different option sets” that we witness in propositions (1*), (2*), and (3*).

From the perspective of practical reason, failures of transitivity within an option set would be much the worse problem, since it could imply contradictory instructions and paralyze choice. This is clearest in cases where the outcomes within an option set are ranked in a cyclical fashion (as in the Up-Down Case). In that case, for each outcome that we could bring about, there is a better outcome that we could also bring about. But even in the our betterness ranking within an option set is not cyclical but merely non-transitive, this could imply contradictory instructions: (1’) suggests that, if I aim to do what will produce the best outcome, I have more reason to bring about B than \( A+ \), whereas (2’) and (3’) together suggest the opposite. We therefore have strong reason to want to deny that the outcomes within an option set can, in an intelligible sense, be non-transitively ordered.22

---

22 By contrast, non-transitivity across different option sets seems far more innocuous from the point of view of practical reason. Since, by assumption, we are not talking about the same context of choice in (1*), (2*), and (3*), this combination of claims cannot imply contradictory instructions or lead to rational dilemmas.
Fortunately, I submit, we are not forced to this conclusion. As the arguments of the following sections will establish, in a three-possible case proposition (3’) is false.

4. Doubts about Independence of Irrelevant Alternatives

Parfit, of course, does assume that his three judgments about the two-possible cases would all carry over to the three-possible case. This is because he accepts an idea often referred to as the “Independence of Irrelevant Alternatives”. Here is the version of this principle that I shall have in mind in what follows:

The Independence of Irrelevant Alternatives (IIA): If outcome $O_1$ is better than outcome $O_2$ when the option set is $S = \{O_1, O_2, \ldots, O_n\}$, then $O_1$ must still be better than $O_2$ when the option set is contracted to $T$, where $T$ is a subset of $S$ that contains $O_1$ and $O_2$. Likewise, if $O_1$ is better than $O_2$ when the option set is $T$, then $O_1$ must be better than $O_2$ when the option set is expanded to $S$.

To some philosophers this principle seems a basic verity about the nature or goodness, almost as unassailable as the claim that the “better than”-relation is transitive. Parfit is one of them. In Reasons and Persons, he writes: “The relative goodness of two outcomes cannot depend on whether a third outcome, that will never happen, might have happened.” If the Independence of Irrelevant Alternatives Principle is correct, then how $A+$ compares evaluatively to $A$ could not depend on whether we are in a 2-possible case, in which these are the only two possible outcomes, or a 3-possible case, where outcome $B$ was also in the feasible set. Hence, any judgement that we make about the relative goodness of $A$ and $A+$ in a two-possible case would necessarily carry over to the three-possible case.

However, while Parfit finds the Independence of Irrelevant Alternatives principle intuitively compelling, I think it is far from secure. To help us see this, it will pay to approach the question obliquely. Instead of asking straightaway whether the presence of $B$ in the feasible set could affect the relative goodness of $A+$ compared to $A$, let us investigate first whether the

---


24 Parfit, Reasons and Persons, p. 429.
presence or absence of B from the feasible set could make a difference to the moral permissibility of choosing A+.

To this question, the answer is emphatically ‘yes’. In a two-possible choice, where the only alternative to A+ is A, the choice of A+ is morally innocuous. Admittedly, opting for A+ creates an outcome in which there is inequality: the New People are worse-off than the A-People, and this fact may be regrettable in itself. Yet, at the same time, their lives are well worth living, and moreover they are the best lives that the New People could feasibly have. It is hard to see how, under these circumstances, creating the New People by choosing A+ could be wrong.

The permissibility of choosing A+ changes entirely if option B is added to the feasible set. It can now no longer be said to the New People that, though they are significantly less well off than the A-People, their lives are going as well as they could have gone. Given that outcome B is also available in this scenario, this is no longer true. Unlike in a two-possible choice between A and A+, there exists a way of avoiding the inequality between the A-People and the New People other than by preventing the New People from coming into existence.

Moreover, there is no good justification for choosing A+ rather than B. Compared to A+, we argued above, B is better on egalitarian, on utilitarian, and on prioritarian grounds. Hence, in the 3-possible case, the New People have an unanswered complaint if A+ obtains and would be victims of injustice, whereas neither is true if the feasible set contained only A and A+. In a 3-possible choice, while it continues to be permissible to choose A, if we are going to bring about an outcome in which the New People exist, our only permissible course of action is to choose B.

The sensitivity of normative properties such as “is unjust” or “is impermissible” on the context of choice is a common phenomenon in ethics. Consider another example:

**Employment:**

A multi-national corporation is considering whether to open a new factory in a developing country, with high rates of poverty and unemployment. It must decide between three options:

1. Don’t hire anyone.
2. Hire 1000 workers, who have very little bargaining power, and pay them very little (though still enough to make working for the corporation at this wage worth their while).
3. Hire the same 1000 workers and pay them a decent wage.
Many would judge that, while (1) and (3) are both morally permissible, option (2) is impermissible, on account of being exploitative.

But what makes option (2) exploitative, or at least wrongfully exploitative? I submit that it is the presence of option (3) in the feasible set. To exploit someone, very roughly, is to take unfair advantage of their vulnerability or weak bargaining position, usually by extracting a bigger surplus than you would receive in a fair transaction. This is the case when the corporation could be paying its workers a decent wage, but uses their weak bargaining position to instead pay them much less than that.

Suppose, by contrast, that the feasible set is different: the world market for the good being produced by the corporation is so competitive that only by paying its workers a very low wage can the company be economically viable. The option of paying a decent wage is simply not available. The feasible set contains only options (1) and (2). I claim that, while it would still, of course, be morally regrettable that the company isn’t able to pay its workers a decent wage, option (2) would, under these circumstances not constitute wrongful exploitation.

My strategy in what follows will be to argue that this sensitivity of normative properties on the context of choice has axiological implications. It seems to me not at all far-fetched that whether an outcome constitutes wrongful exploitation or not might bear, not just on the question whether it is morally permissible to bring it about, but on its goodness. Likewise, I will make the case that, since in Parfit’s Mere Addition Case the presence or absence of option B alters whether A+ is unjust, this eo ipso affects the goodness of A+ compared to A. If this is so, then the Independence of Irrelevant Alternatives principle ought to be rejected and the way is clear to a solution of the Mere Addition Paradox.

5. A Digression: The Fine-Grained Individuation Gambit

It might be interjected that I am misapplying the Independence of Irrelevant Alternatives principle. The principle states that how two outcomes O₁ and O₂ compare in terms of relative goodness cannot depend on whether some third outcome O₃ was also possible, or not. This presupposes that the two

---

outcomes O₁ and O₂ are, in the relevant sense, the *same* outcomes, whether or not O₃ is in the option set.

But it is unclear that this is the case, both in the Mere Addition Paradox and in the Employment case I just discussed. In both instances I suggested that the presence or absence of the third outcome makes a difference to the normatively significant properties of the second (whether it is unjust; whether it is exploitative). This may suggest that outcomes need to be individuated in a more *fine-grained* manner. In the Mere Addition Paradox, A+, when chosen from option set T = {A, A+} and A+, when chosen from option set S = {A, A+, B} are not, in the relevant sense, the *same* outcome.²⁶

I think that, while this intervention is not without merit, it fails to rebut my challenge to Parfit’s assumption that our judgments about the three two-possible cases should all carry over to the three-possible case. This is so for two reasons:

First, recall the dialectical context. It is *Parfit* who needs to invoke the Independence of Irrelevant Alternatives principle in order to make the case that A+ is not worse than A – not just in the 2-possible case (where this claim is supported by his response to the Objection from Inequality) but in the 3-possible case as well (where his response to the Objection from Inequality does not get a purchase). So if the Independence of Irrelevant Alternatives principle could not, in fact, be applied across these two contexts of choice, because A+ isn’t, in the relevant sense, the *same* outcome across these two contexts of evaluation, then Parfit’s argument in the Mere Addition Paradox fails.

Secondly, and more generally, it seems to me that while the strategy of individuating options more fine-grainedly (out of a recognition that the context of choice may alter their morally significant properties) may help to preserve the ‘letter’ of the Independence of Irrelevant Alternatives principle, it already *concedes* the fundamental philosophical point, namely that the context of choice in which a pairwise evaluative comparison is being made is often of crucial importance in making this comparison. The *intended* upshot of the Independence of Irrelevant Alternatives principle is that the choice set in which outcomes occur does not, typically, matter to their relative goodness. However, by resorting to the fine-grained

---

²⁶ For a defense of the idea that putative counterexamples to the Independence of Irrelevant Alternatives Principle can be dealt with by individuating the relevant outcomes or alternatives in a more fine-grained manner, see Michael Neumann, ‘Choosing and Describing: Sen and the Irrelevance of Independence [sic] Alternatives’, *Theory and Decision* 63 (2007), pp. 79-94.
individuation strategy, we accept that context does matter. We accept that, in order to know how two outcomes compare, we often need information about what other outcomes are in the option set. Indeed, we can only know which two outcomes it is that we are comparing in the first place, and what their normatively relevant properties are, by considering facts about the choice set in which these outcomes are situated.

So, it seems that the fine-grained individuation strategy can only “immunize” the Independence of Irrelevant Alternatives principle against the problem cases I am considering by already conceding most of what is at issue, namely that the wider context of choice often matters in evaluating the relative goodness of two outcomes.

6. Why Injustice Makes an Outcome Worse

Here is where we stand. I have argued that

(1) The presence of B in the feasible set gives outcome A+ a property in the 3-possible case – that of being unjust – which it does not possess in the 2-possible case.

And I have already hinted that I believe

(2) The fact that A+ is unjust in the 3-possible case makes it a worse outcome, all else equal, than A+ when the only alternative is A.

If this second claim were also true, it would provide an explanation for how it could be that

(3) While A+ is not worse than A in a 2-possible case, A+ is worse than A in the three-possible cases.

So (2) is the claim that we now must examine.

One source of resistance to (2) might stem from the following line of reasoning: The fact that A+ is unjust when chosen in the three-possible case, whereas it isn’t when chosen in the two-possible case, may certainly make a difference to how one ought, morally, to choose in either case. Agents who bring about A+ in the former case may be acting wrongly, whereas agents who bring about A+ in the 2-possible case may have done nothing wrong.
But these are claims about the difference that the feasible alternatives make to the deontic question whether it is permissible to choose A+.

What we are interested in, however, is an axiological question: do things go worse if A+ obtains in a 3-possible case than if A+ obtains in a 2-possible case? Or, to put the issue a different way: supposing we know that A+ obtains, do we have reason to hope that we are in a 2-possible rather than a 3-possible case, because then A+ would not be unjust, and therefore a better outcome?

By assumption, there is no difference in the levels of well-being that people enjoy in A+ in the 3-option and in the 2-option case. But if the injustice of an outcome doesn’t make things worse for anyone, then why should the presence of injustice matter in itself to the evaluation of the outcome? It is one thing, we might say, to avoid the injustice of A+ by doing B instead. That would actually make things better for the victims of injustice. But how could it be thought to make a difference if the only reason why A+ does not contain injustice is because B, the morally superior alternative, is not feasible?

This skeptical worry would be well-taken if we thought that our only reason for caring whether there is injustice or wrongdoing was out of a concern for the bad effects of injustice of wrongdoing, in terms of people’s wellbeing. But I submit that this is not what we really believe. We can bring this out with what I call the Argument from the News-Value of Wrongdoing.

Consider the following case:

**Bystanders:** You read in the newspaper of a young child who drowned in a shallow pond. You surmise from the story that the child must have stumbled into the pond while unattended, with no-one present to rescue it. However, as you go on reading, you learn to your horror that your initial assumption was mistaken. There were, in fact, a number of bystanders, each of whom could have easily rescued the child, but who instead stood by and did nothing.

If you are like me, finding out that you are a scenario which contained two feasible outcomes, namely

(1) The child drowns.
(2) Bystanders rescue the child.

instead of a scenario in which the latter outcome wasn’t possible, is bad news. Whereas the death of a young child is a tragedy in any event, learning that
what happened was that people stood by and watched the child drown is horrifying. What has happened here, we are tempted to say, is not just a misfortune but a moral catastrophe.

For a second illustration, return to our earlier Employment Case. Suppose you learn of a company which is operating a factory in a developing country with high rates of poverty and unemployment, and is paying its workers very little for long and hard hours of work. Since this is often the case, you assume that what we have here must be a typical case of economic exploitation: in other words, you assume that the company could be paying its workers a more decent wage, but is taking advantage of their vulnerability and weak bargaining position to pay them a pittance instead. However, you subsequently learn that your assumption was, in fact, mistaken. The workers are not being exploited. The company is operating in an extremely tough international market with profit margins that are razor thin. Even by paying its workers as little as it is, it is barely breaking even. With the best will in the world, the option of paying its employees a higher wage is simply not available, if the company is to remain economically viable.

If you are like me, you are relieved to learn this. What you have found out changes nothing about the fact that the workers are being paid very little for long hours of work. But learning that they are not, after all, the victims of unjust exploitation strikes us as good news in its own right. These intuitions about the news-value of information about the presence (or absence) of injustice or wrongdoing lend credence to the view that we care about whether we, or others, are behaving justly or unjustly, once we prescind from the good or bad effects on people’s wellbeing of our so acting?

The contemporary philosopher who has seen most deeply into this question is T.M. Scanlon. For Scanlon, the requirements of morality are “not just formal imperatives; they are aspects of the positive value of a way of living with others.” Behaving morally towards others, not treating them unjustly, allows us to realize a valuable form of relationship with them.

---

27 Neither Parfit nor Temkin would disagree. In fact, in other contexts, both philosophers themselves invoke a notion of deontic badness, according to which an outcome can be worse, all else equal, because it contains wrongdoing or injustice. See Larry Temkin, Rethinking the Good, Chapter 7.4 and Derek Parfit, On What Matters: Volume 3 (Oxford: Oxford University Press, 2017), Section 180.

28 T.M. Scanlon, What We Owe to Each Other (Cambridge, MA: Belknap, 1998), p. 162.
form of *mutual recognition*. It is constituted by living with others on terms that they could not reasonably reject insofar as they also are motivated by this ideal.

Indeed, Scanlon argues that much of our motivation to be moral stems from the appeal of standing in this valuable relationship of mutual recognition with others:

> [W]hat is particularly moving about charges of injustice and immorality is their implication for our relations with others, our sense of justifiability to or estrangement from them. […] [W]hen we look carefully at the sense of loss occasioned by charges of injustice and immorality we see it as reflecting our awareness of the importance for us of being “in unity with our fellow creatures.”

If Scanlon is right, the reason to avoid wrongness, or injustice, is not reducible to avoiding the bad effects of wrongdoing. Wrongdoing has a significance of its own that goes beyond that of causing suffering, or producing other bad effects. By acting wrongly or unjustly towards others, we violate the terms of a relationship of *mutual regard*. By treating them in ways that they could reasonably reject, we fail to accord them a certain form of respect. This puts ourselves in a very different relation to our fellows.

A society or a group of individuals whose dealings are tainted by injustice thus fails to realize a valuable form of interpersonal relationship. That, I think, gives us grounds for not wanting to relate to one another on unjust terms, quite independently of avoiding the bad effects that injustice often has. That is why, all else equal, we have reason to prefer outcomes that are not marred by injustice – even if the absence of injustice does not, otherwise, make people better off.

If these arguments are sound, we are now in a position to assert

(2) The fact that A+ is unjust in the 3-possible case makes it a *worse outcome*, all else equal, than A+ when the only alternative is A.

Now, by itself this is not enough to establish that

---

29 Scanlon, *What We Owe to Each Other*, p. 163.
(3) While A+ is not worse than A in a two-possible case, A+ is worse than A in the three-possible case.

For it certainly isn’t the case that whenever an outcome is unjust, it is necessarily worse than one an agent could permissibly bring about. There are many counterexamples to that claim. In our earlier Employment case, for instance, a morally motivated observer may well hope that, if the corporation doesn’t employ the workers on fair terms (although it could), then it will at least employ them on exploitative terms, even if doing so is wrong. That way, workers will still be spared from the worst ravages of poverty.  

But the Mere Addition Case seems to me different, for the following reason: In the Employment Case, outcome (2) is clearly better than outcome (1) in one respect – there are people, namely the company’s would-be employees, who will be worse off if outcome (1) is chosen. This may make it the case that, despite the fact that exploitation is a bad-making feature of this outcome, this badness can be outweighed, all things considered, by the fact that people are spared from a worse fate in (2).

Crucially, no such thing is true in the Mere Addition Paradox. If we choose A rather than A+, this will not be worse for anyone who ever lives. Moreover, recall that Parfit’s Mere Addition argument is directed at people like myself who embrace the Intuition of Neutrality, and who therefore believe that merely adding happy lives to an outcome does not make the world go better in any other sense either.

If we accept the Intuition of Neutrality, we therefore think that A+ is in no respect better than A. Hence, if in the 3-possible case the presence of injustice makes A+ worse than A in some respect, this is enough to make A+ worse than A all things considered.

If all this is true, we can claim to have solved the Mere Addition Paradox. In the three-possible case, we can affirm the following transitive ranking over the three options: $A > B > A+$. This blocks the argument towards the Repugnant Conclusion. At the same time, we have seen that this ranking in the 3-possible case is consistent with all three pairwise judgments that Parfit affirms in constructing the Mere Addition Paradox, so long as these are understood as evaluative judgments about two-possible cases.

---

30 For a view of this kind, see Alan Wertheimer’s discussion of ‘mutually beneficial exploitation’ in his Exploitation (Princeton: Princeton University Press, 1996).

31 The same, I maintain, is true in the Up-Down Case.
8. Comparison-Dependent vs Context-Dependent Goodness

Temkin is correct that the Mere Addition Paradox has a crucial lesson to teach us about the nature of our moral ideals: it challenges the Internal Aspects View of Outcome Goodness. We should reject the Internal Aspects View, because some evaluative features of outcomes, such as whether the outcome contains injustice or justified complaints, are not just a function of that outcome’s intrinsic properties.

However, Temkin errs when he suggests that the Mere Addition Paradox gives us grounds to reject the Internal Aspects View in favor of a view on which the goodness of an outcome may be comparison-dependent, differing chameleon-like depending on which outcome we are comparing it to. If this were indeed the case, then how A+ compares to A in terms of inequality wouldn’t differ depending on whether or not B was also in the option set or not. Rather, in both a two-possible and a three-possible choice, A+ would not be worse than A in terms of equality, whereas the inequality in A+ would a bad-making feature of A+ when compared to B. In that case, however, our intransitive judgments in two-way choices would carry over to a three-way choice between A, A+, and B, thus raising the specter of non-transitivity within an option set.

Fortunately, Temkin is mistaken that Parfit’s views about the badness of inequality support the claim that how A+ is to be judged in terms of equality is a function of the alternative it is being compared to, as opposed to the option set as a whole. Our basis for thinking that A+ is not worse than A in a two-possible case is that the inequality between the A-People and the New People could only be eliminated by preventing the New People from coming into existence. But crucially, this rationale no longer applies in a three-possible case. Here, there is a way of avoiding the inequality in A+ that makes the New People better off. Hence, not availing ourselves of this option and producing outcome A+ is an injustice; and this, in turn, makes A+ worse than both outcomes A and B in the three-possible case.

The goodness of A+ is not, as Temkin maintains, essentially comparison-dependent, but rather context-dependent. This, as we saw, supports the view that while the correct pairwise judgments in two-possible choices are intransitive across option sets, in a three-possible choice between A, A+, and B, we obtain a transitive ranking. Temkin’s belief that the Mere Addition Paradox poses a deep challenge to the transitivity of the “all-things-considered better than”-relation thus draws the wrong lesson from this case.
8. Dynamic Choice and Worries about Money-Pumping

There is one more loose end to tie up. Even though I have shown that we should not accept that A, A+, and B are non-transitively ranked in the 3-possible case, I do accept each of Parfit’s pairwise judgments when understood as judgments about 2-possible cases. For this reason, I am committed to accepting a weaker form of non-transitivity, namely non-transitivity across different option sets.

One might worry that this might still lead to problems in contexts of dynamic choice, in particular the possibility of being “money-pumped” in a sequential version of the Up-Down Case. Suppose that I am first presented with a choice between A and A++, then a choice between A++ and B, and finally a choice between B and A. At each step, I must pay a small sum of money if I wish to choose the latter option. The worry is that, given my evaluative judgments in the three 2-possible cases, I should prefer to do so, at each step. Yet the overall effect of making these three trades is to return me to my starting point, only poorer. Surely, the objection goes, there must be something rationally defective with a set of evaluative judgments that would support such a pattern of preference.\(^{32}\)

To put this worry to rest, we must distinguish two versions of the money-pump scenario.\(^{33}\) The first is one where I am aware, in making my choice between A and A++, that if I pay to choose A++, I will subsequently be offered the second choice between A++ and B, and that, if I pay to choose B, I will subsequently be offered a third choice, to pay to choose A over B.

This is a scenario, I claim, in which the evaluative truths about A, A++, and B do not, in fact, support my being money-pumped, and moreover I know this. The key point is that in a situation where I will, sequentially, be offered a choice between all three options, their ethical properties do not differ from those in a 3-option case, in which all three options are simultaneously available to me. In particular, if choosing option A++ would be morally unjustifiable in a 3-option case, because the presence of option B gives the New People an unanswerable complaint of injustice against A++, then the same would be true of choosing A++ and then sticking with it in the

\(^{32}\) The money-pump objection to intransitive preferences was first presented by Donald Davidson, J. McKinsey, and Patrick Suppes (1955) in ‘Outlines of a Formal Theory of Value’, Philosophy of Science 22 (1955), pp. 140-60.

sequential case. The fact that B is not *immediately* available as an alternative to A++ is irrelevant to the question whether it is morally justifiable to remain with A++ once the option of moving to B *has* become available.

Hence, in confronting the initial choice between A and A++, I know that if I pay to choose A++, I can subsequently bring about one of three outcomes, each of which is worse than sticking with A in the first place: (i) I could trade to A++ and then stick with this option. But this is worse than remaining with A, on account of the injustice of A++. (ii) I could trade to A++ and, once given the opportunity, pay to trade to B (as I morally ought to), and then subsequently stick with B. But in that case, I have paid twice over to end up with an option that is worse than A. Finally, (iii) I could pay to go from A to A++, from there to B, and thence back to A. In that case, I have paid three times over, only to end up with the same option that I started with. I conclude that, in a scenario with foresight, a rational agent will not be money-pumped.

Suppose, by contrast, that I do *not* know from the start what my options will be. In deciding whether to pay to choose A++ over A, I am not aware that, once I have made this choice, I will subsequently be offered a choice between A++ and B, and so on. Can I now be money-pumped? Yes, I can, given my evaluative judgements about the three 2-possible cases. The question is whether this need indicate any rational defect on my part.

I submit that it does not. There are situations where even a perfectly rational agent can be led around in a cycle, paying at every step, without this indicating anything problematic about the structure of her underlying preferences or evaluative judgments. Consider the following example:

*Change of Plans:*

A single woman is deciding whether to try and become pregnant. Because of a medical issue, she must resort to in vitro fertilization. In addition, she is informed that due to an abnormality in her genotype, any child she will be able to conceive, now or at any point in the future, will suffer from a congenital malady, Condition X, which will restrict its life expectancy to about 55 years. There is currently no known treatment for Condition X. However, other than its foreshortened life expectancy, her child’s quality of life will be unaffected. The woman has the following preferences:

(1) In the absence of a cure for Condition X, she prefers having a child with Condition X to remaining childless.
(2) If she has a child with Condition X and a cure for this condition becomes available, she believes that she is morally required to provide this cure to her child, even if it very expensive, and so she prefers this.

(3) However, given the exorbitant costs of providing a cure for Condition X, the woman would prefer remaining childless to having a child and paying for the cure.

At time $t_1$, believing correctly that no cure for Condition X is currently available, the woman pays a fertility clinic to start a procedure of in vitro fertilization. Shortly after, at $t_2$, she learns that a cure for Condition X has been found. Believing that it is too late to call off her pregnancy, the woman makes a down payment to sign her future child up for the cure.\textsuperscript{34} Finally, at $t_3$, the woman learns that the procedure of in vitro fertilization has not yet yielded a viable embryo but that it soon will; so she pays a small administrative fee to cancel the procedure and ensure that she remains childless.

The woman has paid to trade in a circle. But I can detect no irrationality in doing so. The woman's preferences, both individually and taken as a set, are reasonable and morally permissible. What makes her pay thrice over only to end up where she began is not some defect in the structure of her preferences, but the fact that, at each but the last step of the cycle, she was operating with incomplete information about what her total option set over time would look like.

9. Conclusion

In this paper, I proposed a new solution to Derek Parfit's Mere Addition Paradox. I argued that the paradox trades on an ambiguity about the context of choice. There is a sense in which all three intuitive judgments about Parfit's case are true, namely as pairwise comparisons in two-possible cases. The air of paradox arises from the assumption that these pairwise judgments carry over to a three-possible case, in which all three outcomes are possible. But this, I argued, is not the case. In a three-possible case, there is a complaint of injustice against option A+ and this, I argued, makes it a

\textsuperscript{34} In this stylized case, assume that killing an embryo, in vitro or in utero, is not legally an option.
worse outcome, not only than B but also than A. This solves the Mere Addition Paradox and blocks the argument towards the Repugnant Conclusion. While there are still many impossibility theorems lurking in the field of population ethics\textsuperscript{35}, the threat of the Mere Addition Paradox is more apparent than real.

\textsuperscript{35} The locus classicus is Gustaf Arrhenius, Population Ethics: The Challenge of Future Generations (ms.).