No Harm, No Foul: A Person-Affecting Population Principle

Brandt van der Gaast – University of Twente[1]

Abstract: Why do we care about building a sustainable economy? About providing aid to developing countries? For one, because these policies can result in a better future. But what makes one future better than another? Ethicists and economists have long studied the question of how the overall value of an outcome depends upon the value for individuals that exist in it. This paper defends a person-affecting answer to this question. Overall well-being, on this approach, is captured in terms of relative differences in well-being (i.e. being benefited or being harmed), instead of in terms of absolute levels of well-being (i.e. being well-off or being poorly off). I formulate a person-affecting population principle and discuss some of its theoretical underpinnings. I show how the theory can accommodate egalitarian intuitions and reply to some objections.

1. Introduction

In some moral dilemmas, it is up to the agent to decide who is harmed and who is not. In others, the agent determines who lives and who dies. But there are also moral dilemmas where the agent's action has an effect on the identities of the people to *ever have existed*. These effects are likely to be indirect, but they are not less real. Two actions can differ in that one leads to an expansion of the population while the other one does not. Two actions can differ in that one leads to a population consisting of certain people, while the other leads to a population of certain other people. How do we decide if one such outcome is better than another?

The goal of this paper is to formulate a population principle that ranks outcomes in terms of their overall value. This principle applies to situations where the outcomes have the same population, but also to situations where they differ in their population. The central question I will address is: How does the overall value of a situation depend on the value for the individuals that exist in it? This philosophical question is also important in certain branches of economics. Welfare economics, for instance, studies so-called social welfare functions: mathematical functions that characterize how social welfare depends on individual welfare (Sen 1970, Bossert and Weymark 2004).

The question of how to rank outcomes in terms of their overall value is of interest to consequentialists, but also to non-consequentialists. For instance, some ethicists endorse mixed views, where the value of an action's consequences is merely one of the many components that constrain the moral rightness of the action. Such a view does not count as pure consequentialism, but still needs an account of overall value. Secondly, the question might be relevant to prudential decision-making. Group decisions about which alternative to realize are usually guided by considerations about the overall value of the different possible alternatives.

The plan for paper is as follows. The next section covers some traditional forms of utilitarianism and their drawbacks. In section three I formulate a person-affecting principle of overall value. Section four covers the issue of equality. In section five and six, I discuss some of the theoretical commitments about the extent to which well-being can be measured. I then move on to some criticisms: section seven covers the transitivity objection and section eight covers the asymmetry objection. Finally, I briefly discuss an alternative person-affecting view in section nine.

[1] This paper has benefited from discussions with Michael McDermott, Peter Valentyne, Fred Feldman.

2. Totalism and Averagism

Jeremy Bentham believed that one ought to maximize "the greatest good for the greatest number." But what is the greatest good for the greatest number? Is it the sum of all individual well-being? Or the average? Neither Bentham nor John Stuart Mill addressed this question in much detail. Henry Sidgwick did consider the question. He endorsed a principle that says to maximize "the product formed by multiplying the numbers of persons living into the amount of average happiness" (1847: 415-6). Sidgwick also anticipated certain questions about population expansion. What does an ethical theory recommend we do, he wondered, given that "we can to some extent influence the number of future human (or sentient) beings[?]" (414).

Utilitarian theories usually consist of two elements: a consequentialist component about the relation between the moral rightness of an action and the overall value of the outcome it results in, and an axiological component that captures the determinants of the overall value of an outcome. The latter component usually consists of an aggregation principle about how individual value contributes to overall value. What is a plausible aggregation principle of overall value?

Some utilitarians endorse an aggregation principle that I will call 'totalism'. On totalism, the overall well-being (or utility) of an outcome is simply the sum of the utilities of all the individuals that exist. Other utilitarians endorse what I will call 'averagism', where the overall utility of an outcome is the average of the utilities of all the people that exist.

Oftentimes, agents find themselves in same-people choice situations. In such situations, none of the agent's available actions change the identities of the people to ever have existed. Other situations are same-*number* choice situations. Here, none of the agent's actions change the number of people to ever have existed. Yet other situations are different-number choice situations. Here, outcomes differ in the number (and, so, the identities) of people to ever have existed.

It is well-known that both totalism and averagism have certain problems when it comes to such different-number choice situations. Indeed, many think that standard totalism and averagism are untenable in light of these. Consider the following:

	first	second
A	6	
B	6	5

In this table, every column corresponds to an individual ('first', 'second') and every row corresponds to an outcome ('A', 'B'). The squares in the table represent the utilities of the different individuals on the different outcomes ('5', '6'). These utilities are understood as lifetime utilities, not as time-period utilities. A blank square means that the individual does not exist in the outcome. In the table above, the first individual has a utility of 6 on both outcomes, whereas the second individual exists only on outcome B and has a utility of 5 there.

Totalism ranks outcome B above A. Intuitively, however, outcome B is not better than A. There does not seem to exist a *prima facie* moral reason for 'adding people to the world'. Many authors consider this a strike against totalism. Jan Narveson, for instance, coined the slogan, "We ought to make people happy, not happy people" (1973: 73). Derek Parfit shares the intuition as well. He writes, "if [a] couple do decide not to have [an] extra child, it would not be clear that they are open to moral criticism" (1982:140).[2]

John Harsanyi is an averagist; her writes that, "every possible social arrangement… [is to be evaluated] in terms of the average utility level likely to result from it" (1975: 45). Averagism also faces a difficulty with different-number situations. Consider again the table above. On averagism, outcome B is worse than A because it has a lower average utility. But intuitively, outcome B is not morally worse than A.

The argument against totalism relies on the intuition that outcome B is not better than A. The argument against averagism relies on the intuition that outcome B is not worse than A. If outcome B is neither better nor worse than A, that means that they are on a par. Adding people to the world, in other words, seems to be *morally neutral*. John Broome calls this 'the intuition of neutral existence' and admits that this intuition "grips one strongly" (2004: v).[3]

The intuition of neutral existence will return throughout this paper, so it will be useful to give it a more precise formulation:

> Neutral Existence: For any situation with two possible outcomes, if the outcomes differ only in that one contains a number of additional individuals, then they are equally good.[4]

3. Person-Affecting Principles

Why is outcome B from the previous example not better than A? A possible answer is: because on outcome B nobody is *benefitted*. Similarly, perhaps outcome B is not worse than A because on B nobody is *harmed*. If we take this tack, the concepts of harming and benefiting will take center stage.

A consequentialist, for instance, might adopt a principle like this:

> No Harm, No Foul: An action that does not harm anyone is not morally wrong.

Whether No Harm, No Foul is plausible or not depends on what exactly we mean by 'harm'. Consider the following definition:

> Person P is *harmed* on outcome A if and only if there is an available outcome B where P is better off than on outcome A.

'Benefit' can be defined in analogous fashion (just replace 'better off' with 'worse off'). Combining the two concepts, we can say: a person is *affected* on an outcome just in case the person is either harmed or benefited. One possible person-affecting view is the view according to which the value of an outcome depends only on the harm and benefit done to individuals that exist on the outcome. Adopted into a

[2] J.J.C. Smart famously endorsed this aspect of totalism. He wrote, "Would you be quite indifferent between (a) a universe containing only one million happy sentient beings, all equally happy, and (b) a universe containing two million happy beings, each neither more or less happy than any in the first universe? Or would you, as a humane and sympathetic person, give a preference to the second universe?" (1961 in Smart/Williams 1973: 27-8). But is a preference for a more populated universe really a moral preference? Jan Narveson thinks not; he writes, "we might prefer… a universe containing people to one that does not contain them…, but is this… a moral preference? It seems to me that it is not" (1967: 72). Jonathan Bennett concurs: "I don't regard [my pro-humanity stance] as part of my morality or, therefore, as a source of moral obligations" (1976: 67).

[3] He does not accept it, though. He has "grudgingly concluded it has to be abandoned" (2004: v).

[4] In footnote six, I formulate Neutral Existence for situations with more than two possible outcomes. In section eight, I consider whether Neutral Existence holds in cases where the additional individuals are very poorly-off.

consequentialist framework, the view is that the moral permissibility of an action depends solely on the harm and benefit it causes.

Our definition of ‘harm’ implies the following: A person can only be harmed on an outcome if he or she exists on the outcome in question and on at least one alternative outcome.[5] In other words, a person cannot be harmed if he exists only on one outcome. Another consequence of our definitions—a harmless one—is that in choice situations with at least three alternatives, a person can simultaneously be harmed and benefited on one alternative.

A number of ethicists are initially drawn to such a person-affecting approach—even ethicists who later abandon it in favor of totalism or some variant thereof. Larry Temkin writes about the person-affecting view that “many think it expresses the *essence* of morality” (1987: 168; italics original). According to Peter Singer the view contains “what is fundamentally sound about utilitarianism” (1976: 84). Even Parfit says that “most of our moral thinking” is in terms of the view (1984: 370). Yet all these people abandon the view because of a number of criticisms to be discussed below.

One option is to work out the person-affecting approach by using a principle like:

> Harm Minimization: An outcome is better than another if and only if it contains less total harm.

The total harm on an outcome is the sum of all the harm done to individuals on that outcome. Harm Minimization is stated in axiological terms, but it can also be incorporated into a consequentialist principle. Then it reads: An action is morally right if and only if it minimizes total harm.

Benefit maximization can be defined analogously:

> Benefit Maximization: An outcome is better than another if and only if it contains more total benefit.

In consequentialist terms: An action is morally right if and only if it maximizes total benefit. For same-person choice situations, Harm Minimization and Benefit Maximization are equivalent. But for many different-person scenarios, the two are not equivalent, as we will see below.

If one wants to respect Neutral Existence, Benefit Maximization is not a useful principle. Consider the table below: Outcome A represents no change to the status quo; outcome B represents the addition of an individual with a utility level of 4; and outcome C the addition of the same individual with a utility level of 6. Benefit Maximization judges outcome C to be the best (for it contains 2 benefit). But intuitively, outcomes A and C are equally good.

	first
A	
B	4
C	6

Harm Minimization, on the other hand, generates the right result: outcomes A and C are both best because they both minimize harm.

To see Harm Minimization in action, consider the following choice situation:

first	second

[5] For better readability, I will use male pronouns to refer to nameless individuals or persons.

A		
B	6	5
C		7
D	5	

Here, outcomes A and C are judged best. Outcomes A and C are the outcomes where harm is minimized. Both B and D are sub-optimal, because outcome B contains 2 units of harm, whereas outcome D contains 1 unit of harm.

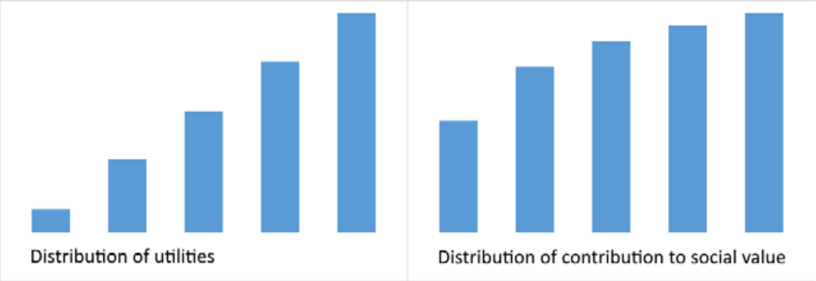
Harm Minimization is the theory that I will adopt and elaborate on in what follows.

4. Inequality Aversion

John Rawls famously endorsed a so-called 'minimax principle'. On minimax, one outcome is better than another just in case the well-being of the worst-off on the former outcome is higher than it is on the latter. In contrast to totalism and averagism, minimax focuses solely on the worst-off. Rawls writes, "Inequalities are permissible when they maximize, or at least all contribute to, the long-term expectations of the least fortunate group in society" (1971: 151).[6]

Minimax is inequality-averse. Averagism and totalism are not and that is one of their drawbacks. A number of authors have addressed this issue by formulating 'generalized utilitarianism' (e.g. Blackorby and Donaldson 1984). The idea here is that the aggregation principle operates on *transformed* individual utilities. Before being aggregated, individual utilities are fed through a transformation function that is strictly increasing and strictly concave. A function is strictly increasing just in case its slope is positive, and a function is strictly concave just in case its slope is decreasing.

Applying a transformation function to individual utilities amounts to a weighing procedure, where increases in utility at the lower end of the spectrum count more heavily towards social value than increases in utility at the higher end of the spectrum. Similarly, decreases at the bottom end subtract more from social value than decreases at the top end. In the diagram below, the left side shows a linear distribution of individual utilities, while the right side shows a concave distribution of their contributions to social value.



One version of generalized utilitarianism is generalized totalism.[7] This view incorporates an aggregation principle that sums transformed individual utilities in order to determine an outcome's overall utility. Another version of generalized utilitarianism is generalized averagism. This view adopts an aggregation

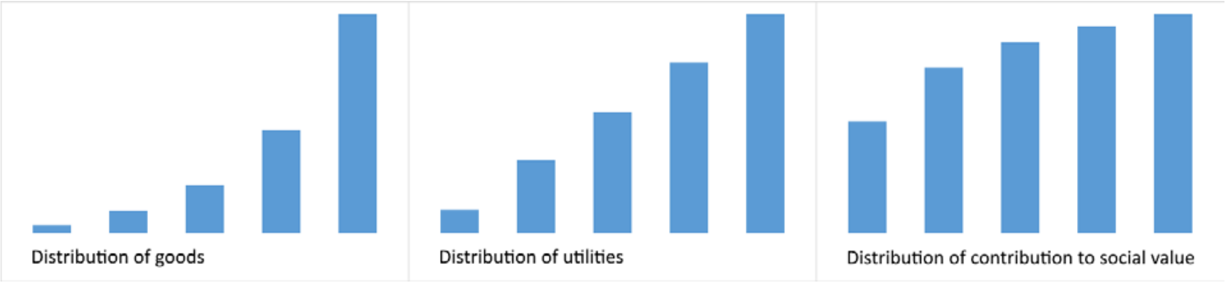
[6] Minimax has the same problem with different-number choice situations as averagism: if the utility of a newly-added person is lower than that of all existing persons, the resulting outcome is worse than the status quo.

[7] See Blackorby and Donaldson 1984. Their version is called 'critical-level generalized utilitarianism'. The view counts as a form of totalism, because a situation's overall value is calculated on the basis of the summation of transformed individual utilities. On their version, the summation procedure applies to transformed utilities *minus* a so-called critical-level. This part of their view is designed to avoid Parfit's repugnant conclusion (1984).

principle that averages transformed individual utilities in order to obtain overall utility. Both views imply that an outcome with a more egalitarian distribution of utilities is better than one with a less egalitarian distribution of utilities, *ceteris paribus*.

It is worth pointing out that using transformed utilities in the calculation of social value is not the same as accepting the diminishing marginal utility of *goods.* The fact that a good (e.g., money) has diminishing marginal utility means that this good's contribution to an individual's utility diminishes as the individual possesses more. The first dollar counts for more than the hundredth dollar, so to speak. This phenomenon is usually captured by applying a strictly increasing, strictly concave transformation function from goods to individual utilities. An increase of goods contributes less to the utility of a wealthy person than does a same-sized increase of goods to the utility of a less wealthy person.

Generalized utilitarianism, in contrast, weighs increases in individual utility for well-off individuals less heavily towards social utility than increases in utility for less well-off individuals. The transformation function is not one from goods to individual utility, but instead one from individual utilities to social utility. If one accepts both the diminishing marginal value of goods and also that social utility depends on transformed individual utilities, then this means that the contribution of goods to social value is adjusted twice. First, a good's contribution to individual utility is adjusted by a principle of diminishing marginal utility, and, second, this contribution to individual utility is then adjusted again in determining its contribution to social utility. See below:



Distribution of goods

Distribution of utilities

Distribution of contribution to social value

Harm Minimization can also be made inequality-averse by using such a function. This requires introducing the notion of *transformed harm*. This can be done as follows: The transformed harm to an individual is the difference between the transformed utility of the individual on the current outcome and the transformed utility of the individual on the outcome where he is best off. Total transformed harm, then, is the sum of all transformed harm. Benefit can be redefined in a similar manner.

To see this modified theory in action, consider the following example:

	first	second
A	4	6
B	7	3

In this choice situation, outcome A contains 3 harm to the first person and outcome B contains 3 harm to the second person. On standard Harm Minimization, the total amounts of harm are the same for both outcomes, so they are on a par.

But after application of the transformation function, the transformed harm done to the first individual on outcome A is *less* than the transformed harm done to the second individual on B. This is because the drop from 7 to 4 occurs at a higher level of utility than the drop from 6 to 3. Since the transformed harm on A is less than that on B, outcome A is better than outcome B. This is in line with our egalitarian intuitions, it seems to me.[8]

A slightly different example of the same phenomenon is the following:[9]

	first	second
A	3	3
B	1	6

On regular Harm Minimization, outcome A (3 units of harm) is worse than outcome B (2 units of harm). But, depending on how strong one's egalitarian intuitions are, outcome A seems as good as B, or even better than B.

If the transformation function has a certain degree of concavity, the transformed harm on A will be equal to that on B. This is because the drop from 3 to 1 occurs at a lower level of utility than the drop from 6 to 3. If this is the case, then both outcomes minimize transformed harm. Both outcomes are then equally good. With a function that is even more concave, the transformed harm on B is *greater* than that on C, and B alone minimizes transformed harm. In that case, outcomes A is better than B.

I will adopt this inequality-averse version of Harm Minimization in the remainder of this paper.

5. Comparisons of Utility

There are different views on how *measurable* well-being or individual utility is. Some views accept only the weaker ordinal measurability, while others also accept the stronger cardinal measurability. Another distinction is that between the weaker intrapersonal comparability and the stronger interpersonal comparability. In this section, I consider which types of measurability and comparability Harm Minimization is committed to.

Suppose we have a complete ordinal ranking of alternatives in terms of how good they are for an individual. This means that any two alternatives are such that, for the individual, the one is better than the other, exactly as good as the other, or worse than the other. Consider a ranking where A > B > C. On ordinal measurability, there is no answer to a question like, 'Is the difference between A and B bigger than the difference between B and C?' To give an analogy: Consider a group of siblings, ordered in a line of increasing age, with the youngest one on the left. From merely looking at the line-up, you know each sibling's rank, but you do not know the exact differences in age among them. The line-up is an ordinal representation of their ages.

A cardinal ordering, on the other hand, is informationally richer. Imagine lining up the siblings in such a way that the distance between any two neighboring siblings represents the gap in age between them.

[8] However, if we change the current same-people choice situation into the following same-number situation, the verdict changes:

	first	second
A	4	
B	7	
C		3
D		6

Here, the harm on outcomes B and D is zero. The transformed benefits on outcome B is larger than the benefit on outcome D. If it is morally right to *either* minimize harm or maximize benefit, then both B and D are permissible. This strikes me also as being supported by intuition.

[9] A variation on this case was suggested to me by Michael McDermott.

This is a cardinal representation. Applied to individual utilities, it means that an individual's utility on one alternative can be represented with a number, and his utility on another alternative with another number. The size of the difference between these two numbers is significant. Cardinal measurability does not attach importance to the choice of unit, or to the choice of zero point. Compare temperature: the Celsius, Fahrenheit and Kelvin scale are cardinally equivalent, even though they differ in their unit and in their zero point.[10]

Ratio-scale measurability is yet stronger. Here, the choice of zero point *does* have significance. On ratio-scale measurability, it is meaningful to speak of absolute quantities of the thing measured. Take the kilometer scale, for instance. It is meaningful to speak of absolute quantities in distance, expressed in a number of kilometers. The kilometer scale is ratio-scale equivalent to the mile scale, because the two coincide on their zero point. If individual utility is ratio-scale measurable, it makes sense to speak of *levels* of utility, and not merely of increases or decreases in utility.

But how can a zero point on the individual utility scale be calibrated? Some utilitarians maintain that a life with positive utility is worth living, whereas a life with negative utility is not worth living. A life with zero utility is a life such that living it is as good for a person as not living it (Broome 2004: 234, 254; Blackorby/Bossert/Donaldson 2005: Ch. 2).

Does it make sense to draw a distinction between lives that are worth living and lives that are not? I do not want to suggest that all lives are worth living, or that no life is worth living.[11] Instead, my point is that it is nigh impossible to get a handle on the distinction between lives worth living and lives not worth living. For instance, I do not think the notion of a life worth living can be straightforwardly connected to the concept of a person's willingness to continue with their life. Accepting a distinction between lives worth continuing and lives not worth continuing does not commit one to accepting a distinction between lives worth living and lives not worth living.[12]

Harm Minimization is a theory that only requires cardinal measurability for individual utilities, and so it does not face the challenge of finding a meaningful way to calibrate a zero point for the individual utility

[10] See Blackorby, Bossert and Donaldson 2005, Chapter 2. In technical terms, the Celsius, Fahrenheit, and Kelvin scale are 'increasing, affine transformations' of each other.

[11] See Benatar 2006 for a defense of this latter claim.

[12] Broome believes that the value of a life (for the person living it) depends upon the value of the stretches of time within that life (2004: Ch. 15). A stretch of time within a life can have zero value, on his view. This happens when the well-being at that stretch of time is at "the level such that a person's continuing to live through an extra period at that level is equally good for her as dying" (235). A life consisting of only such moments, then, is a life of zero personal value. "[A] life that is, throughout, just on the borderline of being worth continuing is, taken as a whole, just on the borderline of being better lived than not lived" (256). It strikes me that this claim relies on an 'intra-life' aggregation principle that is too simplistic to be plausible.

scale. That is an advantage of the theory.[13] In the next section, I will return to views that require ratio-scale measurability. I will question whether we can make meaningful comparisons between absolute levels of utility and relative differences in utility.

There are views that accept intra- but not interpersonal comparability and views that accept both. On mere intrapersonal comparability, the utility of an individual can be compared across times and across alternatives. A person's utility at one time, or on one alternative, can be said to be higher or lower than his utility at some other time, or on some other alternative. On in*ter*personal comparability, one person's utility or well-being can be said to be higher or lower than some *other* person's utility.

Does Harm Minimization require interpersonal comparability? The calculation of harm or benefit to a single person only requires intrapersonal comparability. However, comparing the sizes of harms done to different individuals, or adding such harms in order to obtain total harm, does require interpersonal comparability. Interpersonal comparability is also appealed to in the transformation functions discussed in the previous section. Giving diminished weight to increases in utility at the higher end of the spectrum and increased weight to increases in utility at the lower end of the spectrum requires comparing utilities across people, and so requires interpersonal comparability.[14]

Finally, let me address the issue of summation. HB Maxmin uses a procedure of summation to calculate total amounts of harm and total amounts of benefit. The most famous attack on summation principles in ethics is John Taurek's 1977 article 'Should the Numbers Count?' In this article Taurek discusses trolley-style scenarios and famously argues that there exists no obligation to save the greater number. In these situations, each person deserves your help equally; but it does not follow, he argues, that you ought to save the greater number. If Taurek is correct, then many moral theories that incorporate summation principles are in trouble—including Harm Minimization.[15]

In this paper I argue that Harm Minimization generates moral judgments that in many cases agree with our moral intuitions. This provides us with reasons to accept the theory and so also with reasons in favor of any of the principles that make up the theory. In this section I have claimed that Harm Minimization is committed to cardinal measurability, to intra- and interpersonal measurability, and to a principle of

[13] But why believe that utility is even cardinally measurable? An important argument here is the Von Neumann/Morgenstern theorem from 1947. The key idea in this theorem is that the strength of people's preferences can be figured out by looking at their willingness to take gambles. Consider a standard lottery. For an individual, the value of participating in a lottery depends on the value he attaches to the prizes in the lottery, and the likelihood that he will win those prizes. The Von Neumann/Morgenstern theorem shows the following: On the basis of two sets of facts (viz. the value for the individual of participating in the lottery, and the likelihoods of winning the different prizes), we can infer another set of facts (viz. the value he attaches to the different prizes in the lottery). A simple example can serve as demonstration. Suppose I prefer a burrito to a slice of pizza, and a slice of pizza to a hamburger. Let us now construct a lottery—a simple coin flip—where heads = burrito and tails = hamburger. Would I prefer pizza over playing in this lottery? If so, then the pizza's utility is closer to that of burrito than to that of hamburger. Would I be indifferent? Then pizza's utility is right in between that of burrito and that of hamburger.

[14] See Hammond 1991 for an overview of the issue of interpersonal comparisons of utility.

[15] Unusual versions of the person-affecting approach are still viable, even without any summation principle. For example, consider the view that says to choose the outcome on which the biggest individual harm is the smallest. This version could be called 'Minimax Harm'. This theory still requires intra- and interpersonal comparability, but does not use summation anywhere. Such a view is structurally similar to the well-known decision-theoretic principle of minimax regret. See e.g. Resnik 1987: 28.

summation. So any evidential support in favor of Harm Minimization is indirect evidential support for these commitments as well. As usual, the proof of the pudding is in the eating.

6. Levels vs Differences

Earlier, I quoted Sidgwick who considered the issue of population expansion. Sidgwick wrote, "if we foresee… that an increase in numbers will lead to a decrease in average happiness or *vice-versa*… we ought to weigh the amount of happiness gained by the extra number against the amount lost by the remainder" (1874: 415; italics original).[16] Firstly, 'gained' is not really the right word for Sidgwick to use here, as these extra individuals cannot be said to have any happiness on the alternative where they do not exist. But secondly and more importantly, the totalist calculation that Sidgwick proposes requires ratio-scale measurability and also that absolute levels of well-being are comparable to relative differences in well-being.

But is it part of our moral thinking to make such comparisons? On totalism, the absolute utility of an individual can make an outcome better than an outcome where he does not exist. At the same time, a relative decrease in the utility of an individual can make an outcome worse than an outcome where he is better off. But how do these two types of change stack up against each other? Is it part of our moral thinking to compare absolute levels of well-being with relative differences in well-being?

There are ethicists who want to respect Neutral Existence, but who maintain that in same-*number* choice situations, an outcome with a higher amount of total utility is better than an outcome with a lower amount of total utility. This view amounts to totalism for same-number choice situations. Just like plain totalism, such a view requires that one can meaningfully compare absolute levels of well-being with relative differences in well-being.

Narveson has defended such a view. He writes, "if [you] have a choice of which to produce, [you should] produce the happier one, other things being equal" (1978: 56). Singer has endorsed a similar principle. He writes, "there will be a minimum number of lives being lived [regardless of what we choose], and it is by its effects on the happiness of that number of lives that [an action] should be judged" (1976: 88).

It strikes me that such a view has problematic implications. It runs into trouble, not because it directly conflicts with Neutral Existence, but because it relies on comparisons between absolute levels of individual utility and relative differences in individual utility. For example, consider the following situation:

	extant	first	second
A	5		7.1
B	7	5	

Here, there is one extant person that exists on both outcomes, and there are two newly-added persons, a different one for each outcome. The difference in utility between the two newly-added people is slightly larger than the difference in the extant person's utility on the two outcomes. Standard totalism judges outcome A as better than B. Views like Narveson's and Singer's also judge A to be better than B. Intuitively, however, A is not better than B. It seems that harm to an already-existing person cannot be compensated for by adding a person with higher utility instead of a different person with lower utility.

[16] I am ignoring Sidgwick's "vice-versa". The principles discussed in this paper are formulated in terms of the lifetime utilities of people to ever have existed, and it is impossible for the number of people to ever have existed to decrease.

Someone might bite the bullet and insist that A *is* better than B. This is not a promising strategy, however, because it can lead to a conflict with Neutral Existence. To see this, consider the following situation that has an additional third outcome C:

	extant	first	second
A	5		7.1
B	7	5	
C	5		

Intuitively, it appears that the following things hold: First, outcome A is not better than C (this follows from the intuition of neutral existence). Second, outcome C is not better than B (again, from the intuition of neutral existence). But if A is not better than C, and C is not better than B, it follows that A is not better than B.

The conclusion of this line of reasoning is that outcome A is not better than B relative to the choice situation: {A, B, C}. It seems then that we can also conclude that outcome A is not better than B relative to the choice situation {A, B}. However, this last step in the line of reasoning does not follow without any additional principles. This last step relies on a principle called 'the independence of irrelevant alternatives'. In the next section, I will consider this principle in more detail. But, with the independence principle, Narveson and Singer cannot consistently maintain: i) the intuition of neutral existence, and ii) totalism for same-number choice situations.

Harm Minimization does not run into any of these troubles, because it ranks outcomes purely in terms of harm and benefit. The theory does not require ratio-scale measurability, but only cardinal measurability. This provides the view with two advantages. First, the view is not required to conceptually justify a zero point calibration of the individual utility scale. In the previous section I said this is an advantage of the view, since it seems difficult to establish such a zero point. Second, the view is not committed to counterintuitive comparisons of absolute levels of utility with relative differences in utility.

In this connection it is worth discussing Parfit's non-identity problem (1984: 358). Parfit provides an example where a 14-year-old girl can have a child now or have a child later. If she has a child now, she will not be able to give the child a good start in life; if she waits a significant amount of time, her child's life will turn out much better. Suppose she has her child early. Parfit claims that "it would have been better if this girl had waited, so that she could give to her first child a better start in life" (359). He also maintains that the later child would not be the same individual as the earlier child, due to the different timing of conception.

What does Harm Minimization say about such a case? If the earlier child is the same individual as the later child, then the girl is causing more harm than necessary in having the early child. Assuming everything else to be equal, the outcome with the earlier child is worse. But if the earlier child is not the same individual as the later child, then she is not causing more harm than necessary in having the earlier child, and the two outcomes are equally good. Does the non-identity problem threaten Harm Minimization?

The first thing to point out is that it is not *obvious* that the early child is a different person from the late child. Many views on the metaphysics of personhood have this implication, but it is not a pre-theoretical intuition. The argument contains a somewhat surprising premise (that the two children are not the same person), and a somewhat surprising conclusion (that the outcome with the early child is not worse). As such, the argument is somewhat methodologically suspect. Michael McDermott, for instance, writes, "It is no objection to [the person-affecting theory] that it yields an anti-intuitive conclusion when combined with an anti-intuitive judgment of identity” (1982: 166).

Ignoring this complication, suppose one accepts both that the earlier child is not the same person as the later child, and also that the outcome where the girl has the earlier child is worse than the one where she has the later child. Then one ends up with Singer and Narveson’s view that I just discussed. This view is a form totalism restricted to same-number choice situations. The criticism I presented against that theory still applies: it runs the risk of conflicting with the intuition of neutral existence, when combined with a certain independence principle. To this independence principle I now turn.

7. The Transitivity Objection

Broome presents an example where the agent has the choice of bringing about three outcomes. Outcome A is a continuation of the status quo; B involves the addition of John at 5 utility; and C involves the addition of John at 7 utility:

	John
A	
B	5
C	7

Broome argues: If we compare outcomes A and B pairwise using a person-affecting principle, they are morally on a par, because neither A nor B contains and harm or benefit. Outcomes A and C are also equally good, for the same reason. However, outcome C is better than B, because on B John is worse off than he is on C. Broome writes, “The principle implies, then, that [C] is equally good as [A], [A] is equally as good as [B], but [C] is better than [B]. This is a contradiction. As a matter of logic, the relation ‘equally as good as’ is transitive, and the… principle implies that it is not” (1994: 170).

Many philosophers consider such transitivity arguments to present an insurmountable problem for any person-affecting view. Broome says that they show the view to be “ultimately incoherent” (1994:168) and Parfit that they show the view to have “self-contradictory premises” (1976: 102).

Broome uses *pairwise* comparisons to generate a ranking among A, B and C. But on Harm Minimization discussed in section three, *all* the available alternatives must be taken into account. Harm Minimization judges outcomes A and C both as best, because both contain no harm. Outcome B is less good, because it contains 2 units of harm. I discussed such an example already in section three, where I used to it show that Benefit Maximization generates the wrong result (for it says that outcome C is better than both A and B).

Harm Minimization is an approach to overall value that is sometimes called *deontological*.[17] On the deontological conception, an outcome can be better than another only in relation to a particular choice situation. A different approach is the *axiological* conception. Here, an outcome is better than another just in case the intrinsic overall value of the first is higher than that of the second. Outcomes have such intrinsic values independently of any choice situation. The view that Amartya Sen dubs ‘welfarism’, for instance, is an example of the axiological approach.[18] On welfarism, an outcome’s overall utility is a function only of the utilities of the people that exist in that outcome. Temkin calls the axiological view

[17] E.g. in Tungodden and Vallentyne 2007.

[18] Sen 1979 contains a critical discussion of welfarism. See also Broome 2004: 30-5, 62.

the ‘intrinsic aspect view’: “how good [a] situation is all things considered… will be based solely on the internal features of the situation” (1987: 159).

Harm Minimization rejects a condition known in decision theory and welfare economics as the *independence of irrelevant alternatives* (Sen 1977, 1993). Formulating this principle in terms of the ‘better than’ relation, it reads as follows:

> Independence of Irrelevant Alternatives: If A is better than B relative to a choice set X, then A is also better than B relative to choice set Y, where Y is a proper subset of X.[19]

Harm Minimization does not satisfy this principle. In Broome’s example, outcome A is better than B. But if outcome C is removed from the example, as in the table below, outcome A is no longer better than B. This conflicts with the Independence of Irrelevant Alternatives.

	John
A	
B	5

Is this a drawback? A number of authors have presented examples designed to show that the Independence of Irrelevant Alternatives is too demanding. Sen discusses an example of a guest who is offered cake. Will he take the biggest slice from the plate? Suppose the guest is hungry but also has good manners and therefore takes the second-largest slice. He considers this slice better than all others. Now, what if he had been offered that same plate *minus* the largest slice? Then the slice he actually picked would be the largest slice. But it would no longer be better than all the others. So the removal of an option changes the agent’s ranking of the remaining options. Sen says that this reveals no irrationality on the part of the agent.

Michael Resnik provides another example involving food (1987: 40). A customer is looking over the menu in a shabby-looking restaurant. He sees two items: hamburger and roast duck. The customer fears that the kitchen is not very good, so orders the burger. The waiter then informs the guest that the restaurant also offers sautéed frog legs. The guest now thinks the cook might have skill and goes for the roast duck because he likes duck. Again, this violates the Independence of Irrelevant Alternatives. The addition of an option changes the agent’s rankings: what was previously a sub-optimal choice (the roast duck) now becomes the best choice.

However, these analogies have their limits. Whether the agents in these two examples meet the standards of rationality depends on how their options are described. In Sen’s example, the choiceworthiness of the different slices is not just a function of their size. The goal of the agent is not merely to satisfy his hunger; it is also to make a good impression. And in Resnik’s example, the addition of the third option changes the nature of the first two. As Resnik himself says, “the old acts were *order hamburger at a seedy place*, *order roast duck at the same seedy place*. But the new acts do not include these since you no longer think of the restaurant as seedy” (40; italics original). So Sen and Resnik’s remarks do not settle the issue.

[19] This principle can also be formulated in terms of a choice function, but then its formulation requires two parts. It requires: contraction consistency and expansion consistency. Contraction consistency says: If A is to be chosen from a choice set X, then A is also to be chosen from choice set Y, where Y is a proper subset of X. Expansion consistency says: If A is to be chosen from choice sets X_1, X_2 … X_n, then A is also to be chosen from choice set Y, where Y is the union of X_1, X_2 … X_n (Sen 1993).

Summing up, Harm Minimization is inconsistent with the Independence of Irrelevant Alternatives. Harm Minimization is not a welfarist view where the overall utility of an outcome is merely a product of the utilities of the individuals that exist on the outcome. It is not obvious, however, that disagreeing with the Independence of Irrelevant Alternatives is a big cost to the theory. The burden of proof is on those who insist that a population principle must satisfy this principle.

8. The Asymmetry Objection

Harm Minimization implies Neutral Existence: If a choice situation has two possible outcomes, and if these differ only in that one contains a number of additional individuals, then they are equally good. But what about persons who are very poorly-off? Suppose a woman can have a child whose quality of life is guaranteed to be extremely low. Her alternative course of action is to not have the child. Assume—unrealistically—that everything else is equal. On person-affecting utilitarianism, the woman does not inflict harm if she has the child. Intuitively, however, the outcome where she has the child seems worse than the one where she does not.

	child
A	
B	-4

Many writers on consequentialism and population ethics agree that bringing about outcome B would be morally wrong. Jonathan Bennett writes that "it is wrong to bring into existence someone who will be miserable" (1976: 61). Parfit concurs; he says, "it would be wrong to have the wretched child" (1984: 391).

The standard reply for defenders of a person-affecting approach is to modify the theory. McDermott, for instance, writes, "the following people [are also] relevant…: anyone alive on *one* alternative, if he is miserable on that alternative" (1982: 165; italics original). This is in conflict with Neutral Existence. On this modified theory, non-affected people *can* change the value of an outcome, viz. by lowering its overall value. Other proponents of person-affecting theories that modify the theory in this fashion include Christopher Meacham and Melinda Roberts.[20]

McDermott admits in so many words that this move is *ad hoc*. He writes that the theory "is not *deep* enough. It offers no explanation for the difference… in its treatment of newly-created happy people and newly-created miserable people" (169). Being *ad hoc* is not the final straw for a theory, but it is nevertheless a drawback.

But there is a bigger problem, viz. the problem that I discussed in section six. There, I criticized views that require meaningful comparisons between absolute levels of utility and relative differences in utility. It strikes me that such comparisons are hard to make sense of—regardless of whether these absolute levels represent lives of people who are well-off or lives of people who are very poorly-off.

Consider for instance the choice situation depicted below. How does the harm done to the first individual on outcome B compare to the very low level of well-being of the second individual on outcome A? Which of these detracts more from overall value?

	first	second
A	8	-4
B	4	

[20] Meacham 2012: 263, Roberts 1998: 152 ff.

I submit to have no moral intuitions about situations where relative decreases in well-being (i.e. instances of harm) are weighed against the absolute levels of well-being of people who are very poorly off. To me, this suggests that our judgment about the badness of outcomes where such individuals exist has a different *origin*. What I am suggesting is that these beliefs flow from a different aspect of our moral thinking.

The issue is not that the asymmetry objection relies on the distinction between lives that are and lives that are not worth living. In section five, I claimed that drawing this distinction is not without its difficulties. But everyone has to admit that there are lives that are filled with nothing but misery and suffering. The search is for a theory that can explain why the presence of such lives can detract from an outcome's overall value. The objection cannot simply be dismissed by refusing to accept the distinction between lives that are worth living and lives that are not.

Singer approaches the asymmetry objection from a different angle. He claims, "[I]f for [a certain] reason it is not obligatory to bring a happy person into the world, then by a symmetrical form of reasoning it cannot be wrong to bring a miserable being into the world either" (1976: 93). Singer goes on to say that once such a person comes into existence, there exists a moral obligation to carry out an act of euthanasia on this person. It needs no argument that this is an extreme and also implausible view.

What could be the origin of the intuition that the presence of lives with very low levels of well-being detracts from the overall value of an outcome? Let me briefly and inconclusively sketch how this might be explained. There is a version of consequentialism where not only outcomes have intrinsic value, but the *acts* leading to those outcomes themselves as well. G.E. Moore seems to have held this view. In *Principia Ethica*, he writes, "In asserting that the action is the best thing to do, we assert that *it together* with its consequences presents a greater sum of intrinsic value than any possible alternative." (1903, §17; italics added). A morally right act, on this view, can be an act that "has greater intrinsic value than any alternative [act], whereas both its consequences and those of the alternatives are absolutely devoid either of intrinsic merit or intrinsic demerit" (Ibid.)

Now, if acts can have positive intrinsic value, then perhaps they can have negative intrinsic value as well. For instance, the act of failing to keep a promise could be said to have negative intrinsic value. Now, consider a very poorly-off person that exists on one outcome only. When such a person comes into existence, there exists a moral obligation to improve his life. But often such an obligation cannot be fulfilled. This means that on the outcome where a person with a very low quality of life is created, there will exist unfulfilled obligations. It appears that a situation where there exist unfulfilled obligations is worse than one where they do not exist, *ceteris paribus*.

Whether this suggestion can be developed into a satisfactory reply to the asymmetry objection remains to be seen. But the conclusion from this section is the following. The person-affecting theory can be modified so that the total harm on an outcome not only depends on relative changes in the well-being of people who exist on multiple outcomes, but also on the absolute well-being of very poorly-off people who exist on one outcome only. Against this view, I repeated the criticism that it is difficult to conceptually justify comparisons between relative changes of well-being and absolute amounts of well-being.

9. A Competing View

Let me finally discuss an alternative person-affecting view. This view employs a principle that I will call 'No Avoidable Harm'. Proponents of such a principle include Peter Vallentyne and Roberts (Tungodden and Vallentyne 2007, Roberts 1998). Vallentyne calls the principle 'No Strong Gratuitous Deprivation', while Roberts formulates it in terms of 'wronging'. Since the principle can easily be formulated in terms of the notion of harm, I will stick with that concept. For both of these authors, this principle is merely a small part of their overall theory. Even so, their theory can be criticized by criticizing this particular component.

The principle can be formulated in two steps. First off, there is a principle connecting the notion of moral wrongness to the notion of avoidable harm:

> No Avoidable Harm: An action is morally wrong if it does avoidable harm to at least one person.

Second, there is a definition of 'avoidable harm':

> An action does avoidable harm to a person if it results in an outcome where this person is harmed by being better off on an outcome where nobody is harmed.[21]

Avoidable harm is a special type of harm. Every instance of avoidable harm is an instance of harm, but not every instance of harm is an instance of avoidable harm. In a nutshell: Being avoidably harmed is being harmed by being better off on an outcome where there exists zero total harm.

To provide an illustration of this concept, consider the choice situation below. In this situation, the first individual is harmed on outcome A and the second individual is harmed on outcome B. But only this latter harm to the second individual is avoidable. Outcome B avoidably harms the second individual, because he is harmed by being better off on an outcome where there exists no harm whatsoever (viz. outcome C).

	first	second
A	5	
B	7	5
C		7

Is No Avoidable Harm a plausible principle? I will argue that the principle does not provide enough guidance. The principle is too weak, I claim, because it does not take into account the *size* of harms. Consider the following situation:

	first	second	third	fourth
A	6	3	5	
B	5	6		3

Here, the first and the second individual exist on both outcomes. The first individual is slightly harmed on outcome B and the second individual is more seriously harmed on outcome A. On outcome A, a third individual exists. His level of well-being matches that of the first individual on the outcome where he is worst off. On outcome B, a fourth individual exists and his level of well-being matches that of the second individual on the outcome where he is worst off.

Does there exist avoidable harm on any outcome? No. Both the first and second individual are harmed, but they are not harmed by being better off on an outcome with zero harm. So No Avoidable Harm does

[21] Tungodden and Vallentyne 2007 (Tungodden does not accept No Avoidable Harm). Roberts suggests in a footnote a principle about wronging (1998:63 fn. 48). On this principle, a person is wronged on outcome A if there exists an outcome B such that 1) this person is better off on B, 2) other individuals that exist on both outcomes are at least as well-off on B, and 3) individuals existing only on B are as well-off as is possible. Our principle in the main text amounts to the same thing.

not judge the action leading to A or the action leading to B as morally wrong. It seems to me, however, that outcome A is worse than B. In consequentialist terms, it strikes me that bringing about outcome A is morally wrong.

In order to agree with this intuition, a person-affecting theory that incorporates No Avoidable Harm needs an additional principle. But what principle can that be? The total amount of well-being on both outcomes is the same. The number of people on both outcomes is the same. The distribution of well-being (anonymously considered) on both outcomes is the same. It seems, then, that we need a person-affecting principle to generate the judgment that outcome A is worse than B.

Harm Minimization delivers this verdict. Outcome B is worse than A because the *amount* of harm done on B is larger than on A.

The preceding argument might not win over proponents of No Avoidable Harm. First of all, they might not accept the interpersonal comparability of well-being. A view that does not include interpersonal measurability cannot compare 'sizes' of harm across people in the way that Harm Minimization does. Secondly, it is also possible that someone does not share the moral intuition that outcome A is worse. At this point, the disagreement might result in a stalemate. When offering up a moral intuition, I hope of course to be doing something more than merely spelling out an implication of a certain theory. But intuitions do diverge.

This concludes my discussion of person-affecting consequentialism. Summing up: In this paper, I have formulated a version of a person-affecting approach to population ethics, viz. Harm Minimization. I have shown how the theory can accommodate egalitarian intuitions, I have spelled out its commitments about the extent to which well-being can be measured, and I have tried to show how it can deal with certain arguments from the literature.

But many questions remain. What is the correct explanation of why the existence of very poorly-off individuals can detract from an outcome's overall value? Can the person-affecting principle be extended so that it also takes into account non-human animals? Does the person-affecting theory imply that there is no obligation to ensure the continued existence of the human species? But tackling these and other difficult questions is for another occasion.

Literature

Benatar, David 2006. *Better Never to Have Been: The Harm of Coming into Existence.* Oxford: Oxford University Press.

Bennett, Jonathan. 1978. On Maximizing Happiness. In Sikora, Richard I. and Brian Barry (eds.) 1978. *Obligations to Future Generations*. Philadelphia: Temple University Press.

Blackorby, Charles and David Donaldson 1984. Social Criteria for Evaluating Population Change. *Journal of Public Economics* 25: 13-33.

Blackorby, Charles, Walter Bossert and David Donaldson 2005. *Population Issues in Social Choice Theory, Welfare Economics, and Ethics*. Cambridge: Cambridge University Press.

Boonin, David 2014. *The Non-Identity Problem and the Ethics of Future People.* Oxford: Oxford University Press.

Bossert, Walter and John Weymark 2004. Utility in Social Choice. In Salvador Barberà, Peter J. Hammond, and Christian Seidl (eds.) 2004. *Handbook of Utility Theory, vol. 2: Extensions*. Boston: Kluwer.

Broome, John and Adam Morton 1994. The Value of a Person. *Proceedings of the Aristotelian Society* supp. 68: 167-98.

Broome, John 2004. *Weighing Lives*. Oxford: Oxford University Press.

Hammond, Peter J. 1991. Interpersonal Comparisons of Utility: Why and How They Are and Should Be Made. In John Elster and John E. Roemer 1991. *Interpersonal Comparisons of Well-Being.* Cambridge: Cambridge University Press.

Harsanyi, John C. 1975. Can the Minimax Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory. In John C. Harsanyi 1976. *Essays on Ethics, Social Behavior, and Scientific Explanation*. Dordrecht: D. Reidel Publishing Company.

Narveson, Jan 1967. Utilitarianism and New Generations. *Mind* vol. 76, 301: 62-72.

Narveson, Jan 1973. Moral Problems of Population. *The Monist* vol. 57, 1: 62-86. Reprinted in Sikora, Richard I. and Brian Barry (eds.) 1978. *Obligations to Future Generations*. Philadelphia: Temple University Press.

Narveson, Jan 1978. Future People and Us. In Sikora, Richard I. and Brian Barry (eds.) 1978. *Obligations to Future Generations*. Philadelphia: Temple University Press.

Kavka, Gregory 1981. The Paradox of Future Individuals. *Philosophy and Public Affairs* 11, 2: 93-112.

Meacham, Christopher 2012. Person-Affecting Views and Saturating Counterpart Relations. *Philosophical Studies* 158: 257-87.

McDermott, Michael 1982. Utility and Population. *Philosophical Studies* 42: 163-77.

McMahan, Jeff 1981. Problems of Population Theory. *Ethics* 92: 96-127.

Moore, G.E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

Parfit, Derek 1979. On Doing the Best for Our Children. In Michael D. Bayles (ed.) 1979. *Ethics and Population*. Cambridge: Schenkman Publishing Company.

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

Rawls, John 1971. *A Theory of Justice (Original Edition)*. Cambridge: Harvard University Press.

Resnik, Michael D. 1987. *Choices*. Minneapolis: University of Minnesota Press.

Roberts, Melinda A. 1998. *Child versus Childmaker*. Oxford: Roman and Littlefield Publishers.

Sen, Amartya 1970. *Collective Choice and Social Welfare*. San Francisco: Holden-Day.

Sen, Amartya 1977. Social Choice Theory: A Re-Examination. *Econometrica* Vol. 45, no. 1: 53-89.

Sen, Amartya 1979. Personal Utilities and Public Judgments: Or, What's Wrong With Welfare Economics? *The Economic Journal* 89: 537-58.

Sen, Amartya 1993. Internal Consistency of Choice. *Econometrica* Vol. 61, no. 3: 495-521.

Singer, Peter 1976. A Utilitarian Population Principle. In In Michael D. Bayles (ed.) 1979. *Ethics and Population*. Cambridge: Schenkman Publishing Company.

Smart, J.J.C. and Bernard Williams 1973. *Utilitarianism: For and Against.* Cambridge: Cambridge University Press.

Taurek, John M. 1977. Should the Numbers Count? *Philosophy and Public Affairs* vol. 6, 4: 293-316.

Temkin, Larry 1987. Intransitivity and the Mere Addition Paradox. *Philosophy and Public Affairs* vol. 16, 2: 138-87.

Tungodden, Bertil and Peter Vallentyne 2005. On the Possibility of Paretian Egalitarianism. *Journal of Philosophy* 102: 126-54.

Tungodden, Bertil and Peter Vallentyne, 2007. Person-Affecting Paretian Egalitarianism with Variable Population Size. In John Roemer and Kotaro Suzumera (eds.) 2007. *Intergenerational Equity and Sustainability*. New York: Palgrave Macmillan.

Sigdwick, Henry. 1894. *The Methods of Ethics.* London: MacMillan.

Von Neumann, John and Oskar Morgenstern 1947. *Theory of Games and Economic Behavior (2nd edition).* Princeton: Princeton University Press.