Harm, Benefit, and Non-Identity

Per Algander

# Harm, Benefit, and Non-Identity

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in Geijersalen, Thunbergsvägen 3P, Uppsala, Friday, October 11, 2013 at 10:00 for the degree of Doctor of Philosophy. The examination will be conducted in English.

**Abstract**
Algander, P. 2013. Harm, Benefit, and Non-Identity. Filosofiska institutionen. 198 pp. Uppsala. ISBN 978-91-506-2366-6.

This thesis in an invistigation into the concept of "harm" and its moral relevance. A common view is that an analysis of harm should include a counterfactual condition: an act harms a person iff it makes that person worse off. A common objection to the moral relevance of harm, thus understood, is the non-identity problem.

This thesis criticises the counterfactual condition, argues for an alternative analysis and that harm plays two important normative roles.

The main ground for rejecting the counterfactual condition is that it has unacceptable consequences in cases of overdetermination and pre-emption. Several modifications to the condition are considered but all fail to solve this problem.

According to the alternative analysis to do harm is to perform an act which (1) is responsible for the obtaining of a state of affairs which (2) makes a person's life go worse. It is argued that (1) should be understood in terms of counterfactual dependence. This claim is defended against counterexamples based on redundant causation. An analysis of (2) is also provided using the notion of a well-being function. It is argued that by introducing this notion it is possible to analyse contributive value without making use of counterfactual comparisons and to solve the non-identity problem.

Regarding the normative importance of harm, a popular intuition is that there is an asymmetry in our obligations to future people: that a person would have a life worth living were she to exist is not a reason in favour of creating that person while that a person would have a life not worth living is a reason against creating that person. It is argued that the asymmetry can be classified as a moral option grounded in autonomy. Central to this defence is the suggestion that harm is relevant to understanding autonomy. Autonomy involves partly the freedom to pursue one's own aims as long as one does no harm.

*Keywords:* harm, benefit, population ethics, person affecting view, asymmetry, well-being, reasons, autonomy

*Per Algander, Uppsala University, Department of Philosophy, Ethics and Social Philosophy, Box 627, SE-751 26 Uppsala, Sweden.*



ISBN 978-91-506-2366-6
urn:nbn:se:uu:diva-206059 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-206059)

# Contents

# Acknowledgements

The majority of the time I've been working on this project has been under the supervision of Thomas Anderberg and Gustaf Arrhenius. I am very grateful for their encouragement and the countless occasions when they tried to make me see reason. However, despite their efforts I am certain that this thesis still contains many mistakes. The responsibility for those is all mine.

Sadly, Thomas passed away during the final stages of writing this thesis and I can only hope that he would approve of the final result. At this point, Jens Johansson went beyond the call of duty and volunteered to fill in as an additional supervisor. Jens' thorough comments, disrespect for nonsense and great sense of humour were invaluable during that final push.

I would probably not have come this far without the support of Erik Carlson, Kent Hurtig, Folke Tersman and Jan Österberg. Their willingness to discuss almost any philosophical problem and their never-ending patience sometimes equaled that of my supervisors and I am very grateful for the effort they put into making me improve on this work.

As a Ph.D student I have been very fortunate to have the most excellent colleagues and friends in Niklas Olsson-Yaouzis, Emil Andersson, Tor Freyr, Kamilla Freyr, Olof Pettersson and Rysiek Sliwinski. I would like to thank them all for their moral and intellectual support, as well as for keeping me on my toes with the occasional poke and friendly harassment.

Björn Eriksson read an earlier version of the thesis and I have benefited greatly from his extensive comments. Though his many carefully construed objections were responsible for at least one state of affairs which made my life go worse, I benefited greatly from them all things considered. Katharina Berndt-Rasmussen commented on two occasions on early versions of two chapters. Even though very little remains of those rough drafts in this final version, her comments were very valuable. I would also like to thank Victor Moberger for proof-reading the final version of the thesis and spotting many errors, large and small.

All chapters in this thesis have, in some form, been presented at seminars in Uppsala and Stockholm. I am grateful to all participants at those seminars, especially Sven Danielsson, Hege Dypedokk-Johnsen, Karin Enflo, Karl Ekendahl, Per Ericsson, Lisa Furberg, Mats Ingelström, Magnus Jedenheim-Edling, Sofia Jeppson, Sandra Lindgren, Patricia Mindus, Jonas Olson, Karl Pettersson, Daniel Ramöller, Henrik Rydéhn, Maria Svedberg, Frans Svensson, Olle Torpman and Torbjörn Tännsjö.

A special Thank You to Jesper Holm for allowing me to use “the giant who only wanted to make some friends”.

I would also like to thank Uppsala studentkår, Uppsala University and Helge Ax:son Johnsons stiftelse for their generous financial support.

# Introduction

> Every night and every morn
> Some to misery are born.
> Every morn and every night
> Some are born to sweet delight.
> Some are born to sweet delight,
> Some are born to endless night.
> — William Blake, "Auguries of Innocence"

In 1974 Frank Speck, a victim of the neurological disease neurofibromatosis and father of two daughters with the same condition, decided that it would be best if he was sterilised in order to prevent passing on the disease to any possible future child. A physician was contacted and a vasectomy performed which, the physician assured, would render Mr Speck sterile. Later, however, Mrs Speck became pregnant which prompted the Specks to contact a different physician for an abortion. The abortion was performed and the physician reassured the Specks that the foetus had been successfully aborted despite the fact that Mrs Speck was still pregnant. Some time later Mrs Speck gave birth to a child with neurofibromatosis.

The Specks took the matter to court where one of their claims against the two physicians was on behalf of the child which they had tried to prevent from having. That is, they claimed that the physicians should repair for harm done to the child which would not have existed had either of the physicians not been negligent in their practice. In December 1979, the Pennsylvania Supreme Court refused this claim. According to one of the judges

> [w]hether it is better to have never been born at all rather than to have been born with serious mental defects is a mystery more properly left to the philosophers and theologians, a mystery which would lead us into the field of metaphysics, beyond the realm of our understanding or ability to solve. The law cannot assert a knowledge which can resolve this inscrutable and enigmatic issue. (Speck v. Finegold, 408 A. 2d 496 - Pa: Superior Court 1979).

The judge's puzzlement is certainly understandable. If it would not have been better for the child never to have been born, what is it exactly that the physicians are supposed to repair for? Furthermore, how can they be required to repair for anything if the child would not even have existed had they not been negligent? Here it is important to note that neurofibromatosis is only rarely a life-threatening disease. Most victims of the disease are able to live normal

and by any reasonable standard "good lives". It therefore seems especially dubious to claim that the child would have been better off never existing.

The questions which the story of the Specks raises, and which the judge admitted defeat to, have since the 70-ies been subjected to careful philosophical and jurisprudential scrutiny. On the jurisprudential side, the question whether a child can be *harmed* by being brought into existence has taken center stage. In American law at least, it is necessary in order to establish liability that the defendant (the physicians in this case) has harmed the victim (the child). Many have thought that because the child would not have existed had the defendant not done what he or she actually did it is impossible to establish that the child has been harmed. The child would not have been better off, it is often pointed out, had the defendant not done what he or she did, so there cannot be any harm done.

Philosophers have focussed on similar issues. Harm is a concept which plays a special role in legal practice but also a concept which carries moral weight. However, as Derek Parfit (1984) has argued, when we consider our obligations to future generations, and reproductive choices in particular, it becomes clear that we cannot explain why some actions which we take to be clear cases of wrongdoing are wrong by appealing to harm. In a case like Speck v. Finegold, for example, it seems evident that the child would not have been better off had she not existed so, again, there is no sense in which she has been harmed.

Harm plays a special role in our ordinary moral discourse. For example, harming others is, without a proper justification, wrong. Similarly, that an act would prevent harm is something we take to speak in its favour. While harm is not the only relevant consideration to common-sense morality, it is at the very least an important part. Procreative decisions, and obligations to future generations, pose a special problem however. Intuitively, to do harm is to make someone *worse off* but in procreative decisions the people we are concerned to protect from harm would in some cases, like in Speck v. Finegold, not have existed at all had the putatively harmful act not been performed. How, then, can we in a morally relevant sense harm them?

In this thesis I will argue that for an analysis of harm according to which an act can harm a person even though the person would not have been better off had the act not been performed. The structure of the thesis is as follows. In chapter one I introduce the main principle I will be defending, the Harm Principle. According to this principle, if an act would harm someone then this is a reason against performing that act. I also introduce the main objection to this principle based on harm to future people, the so-called "non-identity problem". I argue that once we unpack the objection to the Harm Principle it is evident that it presupposes a Counterfactual Condition for doing harm: an act harms a person if and only if the person is worse off than she would have been had the act not been performed. This condition for doing harm is the weakest premise in the argument against the Harm Principle and if the principle is to

be saved from the non-identity problem then the most plausible approach is to reject the Counterfactual Condition.

In chapter two I argue that the Counterfactual Condition is not plausible and should be rejected for reasons independent of what we think about harm to future people. The main objection against the Counterfactual Condition is that it has unacceptable consequences in cases of overdetermination and pre-emption. If the effects of a *prima facie* harmful act are overdetermined by another act or event, then it is not the case that the victim of the act would have been better off had the act not been performed. The Counterfactual Condition is therefore not a plausible necessary condition for when an act does harm.

Chapter three discusses one alternative to the Counterfactual Condition, the Non-Comparative View. According to the Non-Comparative View harm should be analysed in relation to a "baseline" state such that an act harms a person if and only if it makes the person worse off than the baseline. While the Non-Comparative View has certain advantages over the Counterfactual Condition I argue that the existing formulations are flawed. It is characteristic of doing harm that it is to make a person's life go *worse* in some sense. The Non-Comparative view does not capture this "contributive character" of harm and should therefore be rejected.

In chapters four to six I present an analysis of harm which I call the Minimalist View. According to the Minimalist View, *a* harms *b* if and only if a person performs an act which (1) is responsible for the obtaining of a state of affairs, *S*, and (2) *S* makes *b*'s life go worse. In chapter four I discuss the second condition and argue that the way in which harms make life go worse is captured by the Same-World View. According to this view, a state of affairs makes a person's life go worse if and only if the person's life is worse for her taking the state of affairs into account than the life is not taking the state of affairs into account.

In chapter five I argue that the first condition should be analysed in terms of counterfactual dependence: an act is responsible for a state of affairs if and only if the act makes a difference to the state's obtaining in a salient way. I compare this analysis with the Counterfactual Condition and argue that this analysis of responsibility has more plausible implications in cases of pre-emption and overdetermination. I also argue that analyses of responsibility in terms of counterfactual dependence in general do not imply the Counterfactual Condition.

In chapter six I argue that we should not add any further condition pertaining to intentions, foresight or consent to the Minimalist View. The reasons which suggest that further conditions are needed are not decisive and often involve intuitions about the overall moral assessment of an act and not only whether it does harm or not. In this chapter I also defend the Minimalist View against the objection that it is too far removed from common-sense. I argue that while it does depart from common sense this is only because in ordinary harm discourse we typically do not make a distinction between "total"

(harm all-things-considered) and "partial" harm (harm-in-a-way). The Minimalist View is an analysis of partial harm, not total, and the fact that the Minimalist View does not account for intuitions about total harm should therefore not count as evidence against it.

In chapters seven and eight I argue that the Minimalist View enables us to save the intuition that the addition of a person to the world with a life worth living can not be required by morality, whereas adding a person with a life not worth living can be forbidden. In chapter seven I argue that it is difficult to reconcile this intuition, sometimes called "the asymmetry" or "the intuition of neutrality", with other solutions to the non-identity problem.

In chapter eight I defend the claim that the asymmetry can be explained by appealing to the Harm Principle and an analogous Principle of Beneficence. This defence of the asymmetry rests on a distinction between two kinds of benefits; benefits that provide only compensating reasons and benefits that also provide requiring reasons. The way to make sense of this distinction, I argue, is to endorse a Principle of Permissibility which states that agents are allowed to perform acts which are not favoured by the balance of reasons if the act does no harm. The Principle of Permissibility captures the importance of respecting people's autonomy and is a way of grounding the claim that benefits to contingent future persons do not provide requiring reasons. Finally, in chapter nine I discuss some of the consequences and possible applications of the Minimalist View of harm and the defence of the asymmetry.

# 1. The non-identity problem

When considering how the present generation can affect future generations it is common among philosophers to distinguish between three different kinds of cases. First, an act may affect how well off future people will be but whatever we do the same people will exist. Call these *same-people cases*. Second, some acts also affect the *number* of people who will exist in the future. When choosing between different courses of action it might be the case that one alternative will lead to there being more or less people in the world than some other alternative. Cases of this kind will be referred to as *different-number cases*. Thirdly, some acts affect the *identity*, but not the number, of those who will exist in the future. For example, a couple who considers whether to have a child at twenty or at thirty are facing a scenario where their choice will lead to different people existing: having one child at twenty or having a different child at thirty.[1] Call cases of this kind *same-number cases*.

In this chapter I will discuss a special problem which arises in same-number cases: the non-identity problem. What the non-identity problem shows is that when an act affects the identity of future people then it is possible that what is intuitively an impermissible choice is not worse for anyone.

The non-identity problem is a problem for many normative theories, so-called person-affecting theories, and theories which appeal to harm in particular. To harm a person, it is often claimed, is to make that person worse off than she would otherwise have been. What the non-identity problem shows is then that we cannot appeal to harm in order to explain why certain identity-affecting acts are impermissible.

The main aims of this chapter are to clarify the non-identity problem and to formulate an argument against the moral relevance of harm. I will also argue that many attempts to solve the problem without appealing to harm are either unsuccessful or comes with counter-intuitive consequences. Finally, I will evaluate the premises in the argument against the moral relevance of harm. I will suggest that the weakest premise is the claim that an act harms a person if and only if it makes the person worse off than she would otherwise have been.

[1] Parfit (1984, p. 351–355) argues for this claim, as does Kavka (1982). It has been claimed by Roberts (2003*a*) that at least Kavka exaggerates how common these cases are. For my purposes it is not necessary that different-people cases are very common, only that they exist.

## 1.1 The problem

In *Reasons and Persons*, Derek Parfit argues that the fact that we can affect the identity of future people presents us with a problem which he dubs "the non-identity problem". Consider the following case, *The Young Girl's Choice*:

> This girl chooses to have a child. Because she is so young, she gives her child a bad start in life. Though this will have bad effects throughout the child's life, his life will, predictably, be worth living. If this girl had waited for several years, she would have had a different child, to whom she would have given a better start in life. (Parfit 1984, p. 358).

In this case it seems clear that the girl ought to postpone her pregnancy. According to common-sense this might be for several reasons but one way of justifying this conclusion would go as follows. If the girl does not postpone her pregnancy then she would harm her child by causing the child to have a bad start in life. Postponing her pregnancy does not involve any significant cost which could justify imposing this harm on the child and there are no other relevant benefits of not postponing pregnancy. Therefore, she ought to wait.

This piece of common-sense reasoning involves a number of considerations. One important normative principle which it relies on is the following:

> *The Harm Principle*: if an act would harm someone then this is a reason against performing that act.

What Parfit and others[2] have pointed out is that it is not evident that the girl can be said to harm her child if she does not wait. The problem with appealing to the Harm Principle, it is often claimed, is that a necessary condition for doing harm is that one makes the victim worse off. More specifically:

> *The Counterfactual Condition*: An act harms a person if and only if that person is worse off than she would have been had the act not been performed.

Suppose the girl does not wait. Then her child will suffer a bad start in life but all things considered the child's life will be worth living. Can we plausibly say that the child is worse off than she would have been had the girl postponed her pregnancy? There are several reasons why this is not plausible. The first is that the child would not *be* at all if the girl had postponed; the child would not have existed. It is therefore unclear whether it makes sense to say that the child would have been better, or worse, off had the girl acted otherwise. The second reason is that even if it can be better (or worse) for this child to exist rather than not, then the least plausible answer in this case is that existence would be worse than non-existence for the child. The child's life is, after all, *worth*

[2] See for example Kavka (1982) and Feinberg (1986).

*living*, and this seems to rule out that existence would be worse for the child than non-existence. We should therefore say either that existence is neither better nor worse for the child or that existence is better. Saying that the child would be worse off existing, and better off not existing, is the least plausible alternative.

One way to reconcile common-sense with the Counterfactual Condition would be to say that, in a sense, the girl makes her child worse off if we read "her child" as a non-rigid designator.[3] As a non-rigid designator, "her child" does not refer to a particular future person but to the girl's child, whoever that may turn out to be. This solution to the problem is not plausible however because it would make the Harm Principle very unintuitive. Harming people, in the sense suggested, is simply not something we care about.[4] What this interpretation of the Counterfactual Condition would suggest is that I would make "my neighbour" worse off, and thereby harm him or her, if I were to move to a new apartment where the referent of "my neighbour" changes from a happy person to a depressed one. A further problem for this view is that it is restricted to same number cases. Reading "her child" non-rigidly is of no help when the choice is between having a child with a life not worth living and having no child at all, though it seems very plausible to say that in such cases the child would be harmed by being brought into existence.

The non-identity problem also arises at a larger scale, as Parfit illustrates with the following example:

> *The Resource Policy*. As a community, we must choose whether to deplete or conserve certain kinds of resources. If we choose Depletion, the quality of life over the next two centuries would be slightly higher than it would have been if we had chosen Conservation. But it would later, for many centuries, be much lower than it would have been if we had chosen Conservation. This would be because, at the start of this period, people would have to find alternatives for the resources that we had depleted (Parfit 1984, p. 361–2).

Suppose this is a same-number case. That is, whether we choose Depletion or Conservation will affect the identity but not the number of future persons. We also assume that life will still be worth living for future persons whatever we choose. Intuitively it seems that we ought to choose Conservation. The common-sense reasoning behind this could be that Depletion involves a small gain in the near future at a large cost in the far future. Conservation on the other hand involves a small cost in the near future and a large gain in the far future. Weighing the costs against the gains gives that we ought to choose Conservation. However, the Harm Principle does not support this reasoning. Choosing Depletion is not worse for anyone because those who will exist in the far future if we choose Depletion would not have existed had we chosen

[3] Hare (2007) argues along these lines.

[4] Parfit (1984, p. 359) makes a similar point.

Conservation. That is, the Depletion-people are not worse off than they would have been had we chosen Conservation. Therefore, we cannot appeal to the Harm Principle to explain why we ought to choose Conservation.

It could be argued that the Counterfactual Condition does not rule out that we can harm future groups of people since a group may be the same even if all its members are replaced. For example, a football team seems to be the same team even if the entire roster is changed over time. In the same way one could argue that a generation does not depend on the identities of the people who make up the generation for its identity. It could then be argued that we do make future *generations* worse off than they would otherwise have been by choosing Depletion, even though we do not make any particular person worse off than she would otherwise have been.

A problem for this reply is that, on the face of it, talk about generations is merely a loose way of speaking about a number of individuals who exist during a certain period of time. A generation seems to be merely a collection of people, like the collection of all red-heads for example, and there is intuitively a difference between such "mere collections" of people and other, less disjoint, groups like a football team. In the latter case it makes sense to say that the group, the football team, is something in addition to the current members of the team. A generation, however, is not a group of this sort.

However, even if we agree that generations are groups in a more robust sense, like football teams for example, a version of the objection raised above against the "non-rigid" response to the case of the young girl applies here as well. I do not harm the group consisting of my neighbours merely by changing the composition of this group from a happy one to a less happy one.

What about different-number cases? It is easy to see that they too present a problem for the Harm Principle since they are, by definition, also cases where different people will exist depending on what we do. For example, suppose that the young girl's choice is between having a child while she is very young or to never have a child. In this version of the case, the Counterfactual Condition does not imply that the girl would harm her child if she decides to have a child for the same reasons as in the same-number version. Any objection against the girl's choice to have a child could therefore not be based on the Harm Principle. It should be noted, however, that the Harm Principle is not the only view which has trouble with different-number cases. These cases raise a host of further problems, such as how to weigh the number of lives lived against the quality of those lives.[5] However, if the Harm Principle cannot account for our intuitions in same-number cases then this is sufficient reason to reject it. The main focus of this thesis will therefore be same-number cases, even though I will consider different-number cases to some extent in chapters seven and eight.

[5] See Arrhenius (2000) for a rather pessimistic conclusion regarding the possibility of finding a moral theory which can solve the many problems connected with different-number cases.

The non-identity problem can be summed up as follows. An act harms a person if and only if that person is worse off than she would have been had the act not been performed. In cases like the Resource Policy it cannot be true that we make people worse off by choosing Depletion. Therefore, it cannot be true that people are harmed in cases like these. However, we clearly ought to choose Conservation over Depletion. Therefore, the reason why we ought to choose Conservation cannot be that we would harm someone and we need to look to some other principle than the Harm Principle to explain why we ought to choose Conservation rather than depletion, or why the young girl ought to postpone her pregnancy.

This argument only establishes that the Harm Principle cannot account for what we ought to do in same-number cases. One could, theoretically at least, hold that the Harm Principle is not intended for same-number cases and that it is therefore not an objection at all that we cannot explain why the young girl should postpone her pregnancy by appealing to harm. However, restricting the principle in this way to same-people cases seems *ad hoc*. If harm is relevant at all then one would expect that the Harm Principle could explain what we ought to do all kinds of cases, not just same-people cases. What we are looking for is not one set of principles for same-people cases and *another* set of principles for same-number cases. Rather, we think that there is one set of principles which accounts for what one ought to do in same-people cases and in same-number cases. The mere fact that a case like the young girl or the Resource Policy are same-number cases, that there is "non-identity", does not seem sufficient to think that different moral principles apply to them.

The claim that identity does not make a difference is however a crucial premise if the non-identity problem is to have any bearing on the plausibility of the Harm Principle. If one were to claim that identity makes a difference then there is no difficulty with combining the non-identity problem and the Harm Principle because the latter might only be one normative principle among many. We therefore need to consider whether identity makes a difference.

## 1.2 Does identity matter?

The argument against appealing to the Harm Principle in order to explain what we ought to do in same-number cases relies on the claim that identity does not make a difference morally. That is, we are not justified in making different moral judgements regarding same-people cases and same-number cases, other things being equal. Parfit (1984, p. 367) calls this view *the no-difference view* and uses the following example to illustrate it:

> *The Medical Programmes*. Suppose there are two medical programmes, *A* and *B*, but that there is only funding for one. We therefore have to decide which should be cancelled. Programme *A* would treat pregnant women who have a

> certain condition. If this condition is not treated it would cause the women's children to be handicapped. This handicap would not be so severe as to make life not worth living. Programme *B* would warn women not to become pregnant during unfavourable circumstances, for example while taking some special medication. If a woman were to become pregnant during these unfavourable circumstances, then this would cause the child to have the same handicap as in programme *A*. A woman who postpones her pregnancy would conceive a different child than if she had not postponed.

Let us assume that the two programmes would be equally effective. Whatever we choose, the result would be that a certain number of children are born without the handicap. If you think that one of the programmes is more worthwhile than the other, then you do not accept the no-difference view. The only difference between *A* and *B* is that programme *A* is a same-people case while *B* is a same-number case. According to the no-difference view this is not enough to judge them differently. The fact that in *A* we are making people better off than they would otherwise have been while in *B* we are merely seeing to it that different people are born is not a morally relevant difference.[6]

The Harm Principle, coupled with the Counterfactual Condition, strongly suggests that we should favour programme *A* in this case. We are not making anyone worse off by cancelling programme *B* and therefore not harming anyone by cancelling it. If we were to cancel programme *A* however we would be making future children worse off.[7]

It could be objected that it does not follow from the assumption that *A* and *B* are equally worthwhile that the no-difference view is true. One possibil-

[6] The no-difference view as it is understood here is similar to the following axiological principle suggested by Arrhenius (2009, p. 290):

> *Impartiality*: If there is a one-to-one correspondence from outcome *A* to outcome *B* such that every person in *A* has the same welfare as their counterpart in *B*, then *A* and *B* are equally good.

Unlike this principle, the no-difference view as understood here does not necessarily rely on *A* and *B* being equally *good*. Rather, the claim is that they are equally worthwhile. On one view, of course, what *makes A* and *B* equally worthwhile is that they are equally good. However, it is preferable at this stage to leave it open exactly what it is that makes the programmes worthwhile. For example, if the Harm Principle is supposed to explain why the programmes are equally worthwhile then such an explanation will be in terms of the harm the programmes would prevent, not in their respective value.

[7] Some would endorse this conclusion and reject the no-difference view. See for example Buchanan et al. (2001, pp. 249–50). This also seems to be the view of some courts as Foster et al. (2006) argues. I argue against this view below. Foster et al. (2006) also notes that philosophers tend to accept either the no-difference view, or that identity makes some difference but not much. They argue that the English law, in contrast, is based on the claim that identity makes an enormous difference. As I argue below, from a philosophical point of view we have good reason to accept the no-difference view. This does not rule out, however, that there can be pragmatic reasons to formulate laws or policies as if identity makes a difference.

ity is that it does make a difference that in cancelling programme *A* we are making people worse off but that in *B* there is some other feature, which is not present in *A*, which makes *B* more worthwhile than *A* with respect to this additional feature. The claim would be that we cannot derive that *A* would be more worthwhile all-things-considered from the fact that *A* and *B* differ merely in the harm-respect. If this were the case then it could be argued that *A* and *B* are equally worthwhile but for different reasons.

While this is a possibility, it is difficult to see what this further difference could be. As was just mentioned, we should assume that the two programmes are similar in all other relevant respects and that the only difference is with respect to identity. There does not seem to be any feature of programme *B* which is exclusive to this programme, that is, any feature which is morally relevant and which cannot be assumed to be a feature of programme *A*.[8]

It could also be objected that the no-difference view is too strong. Consider the following case:

> *Extension or Addition*. Suppose we can choose between implementing either of two policies; *extension* and *addition*. If Extension is implemented, a number of people who would live for 40 years would live for 80 years in stead. If Addition is implemented, an equal number of people, who would not exist if Extension is implemented, would exist and live for 40 years.[9]

Assume that the resources cannot be split between these two alternatives and that implementing either of them will only affect the life-expectancy of these people and not the quality of their lives. What we are choosing between is whether to *extend* the lives of already existing people (without affecting the quality of these lives) or *adding* as many people to the world (with the same short life-expectancy).

The objection to the no-difference view with respect to this case is that the intuitive thing is to favour Extension and this is *because* choosing Addition would be to make people worse off. However, according to the no-difference view this is not a relevant feature. If we accept the no-difference view we would then, according to this objection, be forced to draw the counter-intuitive conclusion that extension and addition are equally worthwhile.

While it is intuitive to favour extension over addition, it is not clear that it is identity that makes the difference in this case. There are further differences between the Medical Programmes and Extension or Addition which might

[8] Steinbock (2009, pp. 169–71) suggests that one programme is favoured by "impersonal" reasons while the other is favoured by "person-affecting" reasons. According to this suggestion, the same-people programme is worthwhile because it benefits people, while the same-number programme is worthwhile because it reduces the amount of suffering in the world. This would not amount to the programmes being equally worthwhile however because the same-people programme also reduces the amount of suffering in the world. The same-people programme would therefore seem to be superior to the same-number programme.

[9] A similar case is discussed by Arrhenius (2008). See also McMahan (2001).

account for why, in the former example, identity does not seem to matter while in the latter it does. For example, one difference between the two cases is that the choice between the two medical programmes will not affect the number of people who will exist while the choice between extension and addition will. On the basis of this difference it could then be argued that the reason we should favour extension is that we should, other things being equal, favour improving the lives of existing people over adding people with lives worth living to the world.[10]

The no-difference view seems plausible, but in fact we do not need to claim that it is true in order to construe an argument against the Harm Principle. What we have to assume is that the two medical programmes *A* and *B* are equally worthwhile. With this assumption, which is significantly more innocent, we can formulate a condensed version of the argument against the Harm Principle in the following way:

> (1) An act harms a person if and only if that person is worse off than she would have been had the act not been performed (the Counterfactual Condition).
> (2) It is not the case that a person with a life worth living is worse off than she would have been had she not existed at all.
> (3) Therefore, a person who would have a life worth living is not harmed by being created.
> (4) If a person who would have a life worth living is not harmed by being created, and the Harm Principle is true, then it would be more worthwhile to cure people of a handicap (invest in programme *A*) than to prevent this handicap from occurring (invest in programme *B*).
> (5) However, programme *A* and *B* are equally worthwhile (the no-difference view).
> (6) Therefore, the Harm Principle is false.

This argument against the Harm Principle can be generalised into an argument against a whole class of normative principles. That it is formulated in terms of harm is not essential. A popular view is that morality is essentially "person-affecting". To capture this it is often thought that the normative status of an act depends in some rather direct way on whether the act affects people for better or worse.[11] For example, if the young girl should indeed wait then we have to explain this in terms of how her choice affects other people.

It is not entirely clear however how to understand the intuition that morality is essentially person-affecting. According to the "narrow" person-affecting view, one state of affairs is better (worse) than another only if the former is better (worse) *for* someone. Assuming that we ought, other things being equal,

[10] This view, sometimes referred to as "the asymmetry", will be more thoroughly discussed and defended in chapters seven and eight.

[11] The person-affecting view is sometimes understood only as an axiological claim about the relation between impersonal and personal goodness. See for example Parfit (1984, pp. 396–400), Temkin (1987, pp. 166–7) and Holtug (2010, pp. 156–63).

to do what would be best the narrow person-affecting view captures the intuition that the normative status of an act depends on whether the act affects people for better or worse and implies that we ought to perform an act only if the act would be better for someone.

The narrow person-affecting view is of course problematic because it is unclear in what sense the girl's choice *could* be worse for her child. As was noted above the claim that the girl's choice would be worse for her child seems to be the *least* plausible answer since the child would have a life worth living. Any normative principle which is person-affecting in this sense is targeted by argument above and will have to explain why we should choose Conservation over Depletion and why the two medical programmes are equally worthwhile.

Alternatively, the intuition that morality is essentially person-affecting can be understood in a "wide" sense. According to this view, the intuition that morality is "person-affecting" is a view about what has value. Raz, for example, claims that "the explanation and justification of the goodness or badness of anything derives ultimately from its contribution, actual or possible, to human life and its quality" (Raz 1988, p. 194). It might be valuable that people have high welfare, or that their rights are respected, but not that the ecosystem is in balance or that scenery is beautiful. The ecosystem and the beauty of the scenery could be valuable because they contribute to human flourishing but not in themselves. According to the wide person-affecting view, morality is essentially person-affecting in the sens that all values are realised by, or in, persons.[12]

The wide person-affecting view is not threatened by the non-identity problem. If the intuition that morality is essentially person-affecting merely means that all values are realised by, or in, persons then this does not purport to be normative so it does not entail anything about what the young girl ought to do, or whether the two medical programmes are equally worthwhile. What the wide person-affecting view amounts to is rather a constraint on theories which aim to explain why the young girl should wait, or what the worthwhileness of the two medical programmes consists in.

The intuition that morality is essentially person-affecting has been a popular view both among those who propose a more consequentialist theory and those who are more on the deontological side.[13] The non-identity problem calls the narrow version of this approach to ethics into question. Though the wide person-affecting view is not threatened by the non-identity problem, it

[12] See Arrhenius (2009, pp. 291–3). Arrhenius distinguishes between a "human good view" and a "personal good view". According to the former, all goods are realised by humans while according to the latter all impersonal goods (bads) are good (bad) for someone. According to the personal good view, then, one need not hold that all impersonal goods are realised by, or in, humans. Both the personal good view and the human good view are however versions of the wide person-affecting view as it is understood here.

[13] See for example Temkin (1987), Holtug (2003, 2010), Scanlon (1998) and Roberts (2003*b*).

cannot solve it either. The wide view merely imposes a constraint on a solution to the problem but it does not amount to a solution in itself.

## 1.3 Ways of responding to the problem

The non-identity problem presents a theoretical problem for a large class of normative principles. There are a number of ways in which one can respond to this problem. The most common response is to abandon the narrow person-affecting approach to ethics and the intuitions that fuel the person-affecting view. What the non-identity problem shows, many seem to think, is that identity is not a morally relevant difference and that it should go the same way as other distinctions which have been thought to be morally relevant such as gender or race.

An alternative solution is to appeal to what would be best from an impersonal point of view in these cases. It would be *better* if the young girl waits, and if we choose Conservation, and this explains why the young girl should wait and why we should choose Conservation. According to this solution to the problem, people ought, other things being equal, to do what would be best. Proponents of this view can also point out that Parfit has suggested a perfectly good explanation of why it would be better if the young girl waits (i.e., why she ought to wait) and why it would be better to choose Conservation over Depletion (i.e., why we ought to choose Conservation), namely the "same-number quality claim":

> *Q*: if in either of two possible outcomes the same number of people would ever live, it would be worse if those who live are worse off, or have a lower quality of life, than those who would have lived. (Parfit 1984, p. 360).

We can see how *Q* solves the non-identity problem. By appealing to *Q* one can claim that it does not matter whether particular people are better or worse off, what matters is that those who will exist if we were to do something are better off than those who would have existed, had we acted otherwise. In the case of the young girl, what justifies the intuition that she ought to wait is that she ought to do what would be best, other things being equal, and it would be better if she decides to wait because the child she would then have would be better off than the child she would have were she not to wait.[14] Similarly, in the case of Depletion, what justifies the intuition that we ought to choose Conservation is that those who would exist if we were to choose Conservation would be better off than those who would exist if we were to choose Depletion. Conservation is therefore the better policy and the one we ought to choose. *Q* is also consistent with the no-difference view because it does not attach any

[14] Note that a proponent of this solution to the problem need not claim that one ought always to do what would be best. The "other things being equal"-clause is important because it allows for the moral status of an act to depend other factors than goodness.

importance to whether the programmes make future people better or worse off, but only to how well off those who will exist if we were to choose one of the programmes would be compared to how well off those who would exist if we were to choose the other programme would be.

Advocates of this approach agree that the Harm Principle gives the right result as long as we are dealing with same-people cases. However, as soon as we move to same-number cases the Harm Principle gives the wrong result and we should therefore appeal to *Q* instead. This is because if we only consider ordinary cases where what we choose will only affect people's well-being, then *Q* cannot be distinguished by what it prescribes from the Harm Principle. If, for example, the young girl could have had the same child even if she had waited, then both *Q* and the Harm Principle would justify the claim that she ought to wait. Since *Q* gives the right verdict in these cases and in same-number cases it is plausible to say that *Q* replaces the Harm Principle.[15] Furthermore, *Q* is clearly compatible with the wide person-affecting view because *Q* is a formal claim about what makes one outcome better than another, not a claim about what has value.

Extending *Q* to work for different-number cases is more problematic,[16] but such cases raise further difficulties which *Q* was never intended to handle anyway. As I have also mentioned, different-number cases are also a problem for the Harm Principle. That *Q* does not solve these cases is therefore not something which should make us prefer the Harm Principle (or any narrow person-affecting principle) over *Q*.

Proponents of this way of dealing with the problem can then claim that (i) we can solve non-identity problem by appealing to *Q*, and (ii) this solution does not force us to give up anything of significance. Therefore, the non-identity problem can be considered solved.

However, we should not be so quick to accept the claim that we do not have to give up anything of significance by appealing to *Q* in same-number cases. For example, suppose a couple has to decide whether to have a fortunate child now, who would be very well off, or a "normal" child later. In this case *Q* implies that it would be better to have the fortunate child, but it is not obvious that one ought to have the fortunate child. It strikes many as deeply unintuitive and elitistic to deny that it is permissible to have the normal child in such cases. It might be objected that *Q*, an axiological claim, does not suggest anything about what people ought to do. It simply says that it would be better to have the fortunate child and nothing more. While this is certainly true, this defence of *Q* would undermine its effectiveness in dealing with the non-identity problem. The non-identity problem is, as I said above, the problem of explaining why the young girl ought to wait, and why we ought to choose Conservation over Depletion. *Q* was put forward as an answer to that question. If we do not

[15] See Parfit (1984, pp. 370–1).

[16] See Parfit (1984, chs. 17–19) and Arrhenius (2000).

think that there is a connection between what we ought to do and betterness, then $Q$ will not even count as a solution to the non-identity problem.[17]

Furthermore, while appealing to $Q$ is compatible with the wide person-affecting view it is questionable whether this is sufficient to capture the intuition that morality is person-affecting. After all the explanation of why the young girl ought to wait is not because waiting would be better for that child. Rather, the explanation is in impersonal terms. That is, the explanation of why girl ought to wait seems to have little to do with what she does to her child. By appealing to $Q$ what we can say is that the girl should wait, not because she does something bad to the child she would have if she does not wait, but rather because she could do much better for some other child. But, this does not capture very well the intuition that morality is person-affecting. Intuitively, what justifies the claim that the girl would do wrong if she does not wait seems to be the intrinsic nature of the state of affairs which she would then bring about. Perhaps it is not the only reason to object to the girl's choice, $Q$ might be one relevant factor, but it certainly seems to be *a* reason.

This suggest is that the second claim above, that solving the non-identity problem by appealing to $Q$ does not carry any significant costs, is doubtful. In what follows I will discuss three person-affecting solutions to the non-identity problem which have been put forward as alternatives to appealing to $Q$: conditional duties, wronging and harming. There are three things to keep in mind when considering these alternatives. First, these alternatives should actually solve the non-identity problem and give a plausible explanation of why the young girl ought to wait. Second, for these alternatives to actually be superior to $Q$ they must also explain some intuitions which $Q$ does not account for. Finally, an alternative solution should either be consistent with the no-difference view or show why this view is mistaken.

### 1.3.1 Conditional duties

It has been suggested that the non-identity problem can be solved by conceiving of duties to future persons as conditional duties. By a conditional duty I mean a duty whose existence is conditional on certain acts being performed. A promise to $\phi$, for example, creates a duty to $\phi$ which is conditional on the act of promising in this way. A duty to promote the good, on the other hand, is not conditional in this way. If an act promotes the good, then the duty to perform this act does not depend on whether some other act have been performed in the past.

The idea is that duties to future persons are conditional in the sense that we can now make it the case that in the future we will have certain duties to people who will then exist. On this view present persons do not have duties

[17] I will postpone a more careful discussion of these matters to chapter eight. Here it is enough to note that $Q$ together with the assumption that there is a reason to promote the good has certain counter-intuitive consequences which support including some person-affecting consideration.

to future persons in the same way as they have duties to their contemporaries. Rather, we only have duties to people who currently exist but we can make it the case that we will have duties in the future to people who will exist. In the case of the young girl, for example, the reason she ought to postpone her pregnancy is that if she does not then she will create duties for herself which she will not be able to fulfill.

Let us make this view more precise. The view is that what makes it impermissible for the young girl to have her child now is that it would create duties for the girl which she cannot fulfill. One way of understanding this is as follows:

> *The Conditional View-1*: One ought not to perform acts which make it impossible for oneself to fulfill future duties.[18]

This view assumes that there are some set of duties which we have to our contemporaries. Exactly what these duties are is rarely articulated but let's for the sake of argument assume that there is a duty to give people a good start in life. With this assumption it could then be argued that when the girl does not decide to postpone her pregnancy she performs an act which makes it impossible for her to fulfill her future duties, namely the duty to give her child a good start in life.

It certainly seems plausible to say that one should not put oneself in a position where one cannot do what one ought to do, if such cases exist (more on this below). This cannot be the whole story about our obligations to the future, however. Consider a case where we can bring about a disaster in the far future. Presumably we should not bring about disasters in the far future but the Conditional View-1 seems badly equipped to justify this belief because *we* will not exist in the far future. *We* will not have any obligations which *we* cannot fulfill.

In an attempt to make the view more plausible, the conditionalist might suggest the following principle instead:

> *The Conditional View-2*: One ought not to perform acts which makes it impossible for anyone to fulfill their future duties.

By saying that we should not create duties for anyone which they cannot fulfill one can accommodate the intuition that our duties to future persons extend even to the far future where we will no longer exist.

A problem for both versions of the Conditional View is that they rather blatantly violate "ought implies can". The Conditional View is based on the assumption that there are situations where it is my duty to $\phi$ but it is not possible for me to $\phi$. Note that this is not the same claim as the more plausible

[18] Narveson (1973, p. 73), Parsons (2002, p. 145) and Vanderheiden (2006, p. 344) hint at a view of this kind though none of them gives it a very precise formulation.

one that there are situations where the conjunction of all one's duties cannot be satisfied but where each duty is possible to satisfy. Consider for example promises again. The conditionalist might point out that if one promises Black to $\phi$ and White not to $\phi$, then one is in a situation where one cannot fulfill all one's duties. However, this is a case where, presumably, one can $\phi$ and one can refrain from $\phi$-ing. The problem is, of course, that one cannot do both. What the Conditional View suggests is something much more extreme, namely that it could be the case that one ought, or has a duty, to $\phi$ but it is impossible that one $\phi$s. This, I think, is enough reason to reject it.

In order to avoid this problem the conditionalist might suggest that we should reformulate the conditional view once more:

> *The Conditional View-3*: One ought not to perform acts which create duties which someone *will not* fulfill.

According to this version of the conditional view, what is important is that we do not create duties that *will not* be fulfilled rather than duties that *cannot* be fulfilled. With this modification the conditional view is no longer in conflict with "ought implies can" while still accounting for duties to future persons.

While the Conditional View-3 is an improvement, I think there are more general arguments to be had against this approach to the non-identity problem in general. First, appealing to conditional duties does not account for our person-affecting intuitions. According to this approach, the objection to the young girl's choice to have the child has nothing to do with the nature of how this choice will affect her child. Rather, it has to do with something akin to the agent's moral integrity: we ought not to create persons who will have a bad start, neurofibromatosis or what have you because it will make us unable to perform our duties. It is only in an indirect way that our duty to future people depends on the effects of our choices on them. While preserving one's moral integrity might be one factor, it is hardly the decisive one in cases like the ones we are considering.[19]

Second, a question which so far has not been answered is why we have to bring in these conditional duties in the first place. The Conditional View obviously assumes a set of duties; those which are held to contemporaries. Why cannot these be applied across times in a more straightforward fashion? Presumably there will be duties to promote the well-being of others (at least to an extent), not to harm others and so on. The Conditional View does therefore seem committed to explaining why future people do not come under the scope of these duties.

One reason why the Conditional View might seem plausible concerns metaphysical worries about duties to non-present people. The worry is that if future

[19] One consequence of the self-regarding nature of the Conditional View is that it seems rather fetishistic with respect to duties. I will ignore this objection however since there are further, and more serious, objections to the Conditional View.

people do not exist, now, then they cannot be said to exist at all and it is not possible to have any duties towards people who do not exist.[20] I think this worry is exaggerated. Everyone agrees that non-present things do not exist now but this trivial claim is irrelevant to whether acts which affect the future can be impermissible in virtue of their effects. If the consequences of an act *would be* very bad for a future person then it is this fact, not the fact that it *is* very bad for that person, which makes the act impermissible (or at least counts against doing it). What the conditionalist would have to claim is that future facts, such as what the consequences of an act would be, cannot be what makes the act right or wrong since the consequences do not exist when the act is performed. This seems to be inconsistent with what the conditionalist wants to say. The idea is that we should not do certain things because this will in the future cause a situation where people will not do (or will not be able to do) what they ought to do. But this is just to say that the impermissibility of an act depends on future facts; that some people will not do what they ought to do. The metaphysical "rationale" for the conditional view actually undermines it rather than supports it.

### 1.3.2 Wronging

An alternative approach to the non-identity problem is to argue that while the young girl does not harm her child she ought to postpone her pregnancy because if she does not then she *wrongs* her child. The idea is that by affecting people in certain ways one can "wrong" them and that affecting people in this way, that is, wronging them, is what makes an act impermissible. This approach preserves the person-affecting intuition since the reason the girl ought to postpone her pregnancy is because of the effects this choice has on her child. If this view does not face the same problems as the Harm Principle when it comes to the non-identity problem then it would have a considerable advantage since it solves the problem and preserves the person-affecting intuition.

Crucial to this approach is of course what one means by "wronging". A natural way of interpreting "wronging", at least as we tend to use it in ordinary language, is in terms of rights. Rights are, by most definitions, personal and are sometimes thought of in terms of claims or demands. A failure to respect a right is to treat someone, the bearer of the right, in a way which is not in accordance with what one owes the bearer, i.e., the bearer can claim or demand not to be treated in this way.[21]

Rights are very complex things, and they are used in a number of different ways. To simplify the discussion I will adopt the following terminology. Let's

[20] This objection has been raised by for example Macklin (1980) and De George (1980). For a reply, see Elliot (1989).

[21] According to the traditional analysis of rights, claim-rights are merely a species of rights. See Hohfeld (1966). It should however be noted that Hohfeld's analysis is primarily an analysis of legal rights and not moral ones.

say that the *content* of a right is a state of affairs. This content is specified by what the bearer has a right *to*. For example, the content of the right to health is the state of affairs that the person who possesses the right is healthy. Let's also say that a right is *satisfied* if and only if the content obtains. That is, a right to health is satisfied if and only if the bearer of the right is healthy. Negative rights can be treated in the same way. A right not to be tortured is satisfied if and only if the content obtains. That is, the right is satisfied if and only if it is false that the person is tortured. Finally, let's say that rights ought to be satisfied, other things being equal.

This terminology does not tell us anything about the content of rights, i.e., what rights there are. A common view is that rights are connected with people's *interests*: a person has a right if and only if it is in that person's interest that the right is satisfied.[22] Formulating the connection between rights and interests in this way is very vague but it may still serve as a guiding principle when discussing the content of a theory of rights.

With this terminology in place, one way of understanding what it means to wrong a person is to say that *a* wrongs *b* if and only if *a* sees to it that a right possessed by *b* is not satisfied. For example, Jeffrey Reiman holds that "in choosing the negative policies [i.e., Depletion], one has wronged the future people who are negatively affected as a result – even though the alternative is that those people would not have existed at all. Indeed, I contend that in these cases, living people are violating the rights of future people" (Reiman 2007, p. 72). Reiman then argues that the in typical non-identity cases, like the case with the young girl or Depletion, one would fail to satisfy future people's right to "a normal level of functioning" by making the intuitively impermissible choice.

An objection which has been raised against this approach is that it is problematic to talk about rights without bearers.[23] The future people whose rights are supposed to explain how our choices can wrong future people do not exist now and therefore their rights cannot exist now either. But, the objection continues, if the rights which are supposed to explain how we can wrong future people do not exist now then they cannot explain any duties which we have towards future people. This objection has much in common with the kind of reasoning mentioned at the end of the last section where it was claimed that some people are attracted to the conditional approach for metaphysical reasons. As we saw, however, there is nothing metaphysically mysterious about duties to future persons and the same reasoning can be applied here. Future people will have rights when they exist, and choices we make can bring about states of affairs that will satisfy or fail to satisfy these rights. That the bearers of these rights do not exist at the time of the choice, or that these rights cannot be satisfied at that time, is beside the point.

[22] This so called "interest view" of rights goes back at least to Bentham. See also Waldron (1984) and Raz (1984).

[23] See Macklin (1980) and De George (1980).

A more serious objection to rights as a solution to the non-identity problem is that in order for the rights-approach to be an alternative to the Harm Principle it cannot rely on a right not to be harmed to solve the non-identity problem. The rights-approach cannot simply assume that the child would violate the child's life not to be harmed because, as we have seen, the child is not worse off than she would otherwise have been. The rights-approach therefore has to find some other right which the young girl would fail to satisfy by not postponing her pregnancy. However, it is quite unclear what right this would be. *Prima facie*, it is the right not to be harmed which is violated in the case of the young girl. In order to understand the content of a theory of rights, it seems expedient to rely on harm. The rights-approach therefore does not have a clear advantage over the Harm Principle when it comes to solving the non-identity problem.

It might be suggested that the right which the young girl would violate is the child's right to a decent start in life and that this right should not be spelled out in terms of harm. Steinbock (1986), for example, claims that "it is a wrong to the child to be born with such serious handicaps that many very basic interests are doomed in advance, preventing the child from having a minimally decent existence to which all citizens are entitled" (Steinbock 1986, p. 19). It could however very well be questioned whether this right is independent of a right not to be harmed. It is not clear that we should care about "doomed basic interests" unless this also involve harm.

I will not push this objection to Steinbock's view however because there is a more general objection to the rights-approach. If rights are always "in the interest" of their bearers (see above) then the rights-based approach faces much the same problems as the Harm Principle. The right which is supposed to be in the child's interest cannot be satisfied, and it is unclear in what sense it is in the interest of the young girl's child to have such a right to a decent life satisfied. The child cannot have a better start in life and of all the available alternative open to the girl, not postponing is best for this child. The girl is therefore clearly acting in this child's best interest by not postponing her pregnancy.[24] The rights-based approach to the non-identity problem does therefore not avoid the objections which were raised against the Harm Principle above.

In defence of the rights-approach it might be suggested that harm is not prior to rights, or wronging, and that the interest-view of rights does not capture *all* rights. An objection along these line has been raised by Kumar (2003) who argues that harm, understood as a "setback of one's interests", is neither necessary nor sufficient for wronging. Kumar argues for this claim by considering various examples which purport to show this. First, that harm is not necessary for wronging is illustrated by an example of a drunk driver. What makes drinking and driving wrong, Kumar thinks, is that it exposes others to risk and this can be said to wrong those who are exposed to this risk. A drunk

[24] See McMahan (1981, p. 126) for a similar claim regarding whether a rights-based approach could solve the non-identity problem.

driver need not harm anyone for his behavior to be impermissible. Merely exposing a person to a risk is not to harm the person on Kumar's view because "the risk did not in fact blossom into an actual harm, or end up setting back one's interests in any way" (Kumar 2003, p. 103). Harming is therefore not necessary for wronging.

Turning to whether harm is sufficient for wronging, Kumar merely states that "[c]onduct may result in another being harmed without it being a moral violation" (p. 100). What Kumar seems to have in mind here are cases where one is justified to set back a person's interests but which are nevertheless not wrong or even objectionable to an extent. Typical examples of this are cases where the harm is necessary for an equal or greater benefit. Amputating a person's leg in order to save her life seems to be a case where we do harm in a sense but where this harm is justified. Amputating a leg in order to save a life does not constitute a "wrong" done to the person, because of the greater benefit (the person's life is saved). Other examples are cases involving prior agreement. The participants in a professional (and fair) boxing-match, or competitors on a fair market, impose setbacks to each other's interests, but there does not seem to be a moral violation here. Another example: a person who has been injured at a construction site is perhaps not "wronged" if the person had agreed to working at the site and the security measures were adequate.[25]

Kumar's examples are not entirely convincing for a couple of reasons. For instance, Kumar's claim that the ordinary concept of harm does not count exposure to risk as a harm could be taken as an argument in favour of expanding our ordinary concept of harm. Indeed, Kumar's claim that being exposed to a risk of harm is itself not a harm because it is not setting back one's interests could very well be questioned. Kumar's motivation for not saying that risks (can) constitute harm is that nothing actually happens to the victim. But, exposing someone to risk is for something to happen to that person. The risks Kumar refers to are not "free-floating", they have subjects just like broken limbs, heart-failures and so on. It is also not obviously counter-intuitive to hold that it is "in a person's interest" to not be exposed to risks. One could therefore hold that reckless driving under the influence does harm to others because it exposes them to a risk.

Regarding whether harm is sufficient for wronging it could also be questioned whether life-saving amputation is a "genuine harm". We could, for example, distinguish between harm in a wide sense and harm in a narrow, or moral, sense. In the narrow sense one does no harm by performing life-saving surgery, but in a wide sense one does. If we add that harm in the wide sense is not sufficient wronging, while harm in the narrow sense is, then we seem to have accounted for Kumar's counterexamples.

[25] There is, I suppose, a much more complicated story to tell about this example. For instance, mere agreement from the worker does not seem to be enough. We would also require, I think, that the agreement was reasonable, not made under duress and so forth.

I will not push these points further, however, because it would take us too far into the finer details of an analysis of harm; whether harm should be understood as setbacks of interests and whether there is a "morally relevant sense" of harm. These matters will be pursued in greater detail in the following chapters. Instead, let us grant for now that harming is neither necessary nor sufficient for wronging and turn to Kumar's way of characterising this notion.

Kumar characterises wronging as such:

> One person wronging another, then, requires that the wrongdoer has, without adequate excuse or justification, violated certain legitimate expectations with which the wronged party was entitled, in virtue of her value as a person, to have expected her to comply. (Kumar 2003, p.107).

To harm a person is sometimes to "violate a person's legitimate expectations" though not always, and there are other such expectations on Kumar's view which are not related to harm.

A preliminary objection to Kumar's way of characterising wronging is that it is unclear what it implies in same-number cases. In the case of the young girl for example, the child she would have if she decides not to wait cannot reasonably expect any outcome other than a life with a bad start or non-existence. It would be unreasonable of the child to expect any other outcome because those are the only possible outcomes. If that's the case, however, it is not clear in what way the girl would "violate" any legitimate expectations on the child's part. The child, once she exists, certainly would prefer existence over non-existence since her life is worth living. How Kumar expects this notion of wronging to explain why the young girl ought to wait is therefore less than obvious.

It could be argued in Kumar's defence that the girl's child cannot have *epistemically* legitimate expectations on a better start in life but that the child can have *normatively* legitimate expectations on a better start. In other words, the child could object to the girl's choice, not on the grounds that the girl could have done better for her, but on the grounds that she *should* have done better. However, saying that the child can have normatively legitimate expectations to a better start is just to assume what has to be shown; namely that it would be morally objectionable of the girl not to postpone her pregnancy. It is this intuition which the appeal to wronging is supposed to justify but with this reply it is merely assumed.

A further objection to Kumar's view, and the appeal to wronging in general, concerns the Medical Programmes. Recall the structure of the case. If we choose to implement programme *A*, the same people case, then we would make a number of future children better off than they would otherwise have been. If we implement programme *B*, the different people case, we would not be making future people better off than they would otherwise have been, but

we make it so that a number of healthy children are born instead of unhealthy children.

With respect to this case Kumar has the following alternatives: (i) they are not equally worthwhile (ii) the two programmes are equally worthwhile because they prevent wrongs, *not* because they prevent harms, (iii) they are equally worthwhile because they prevent harm *and therefore* wrongs, (iv) they are equally worthwhile because one prevents harm while the other prevents wrongs.

First consider option (i). As before, we should assume that the merits of these two programmes are wholly dependent on their effects on the children's well-being and that other factors (such as desert) are equal. The problem for Kumar is that in the same-people case (programme *A*) we can say, assuming the Counterfactual Condition, that it is worthwhile because it prevents harm, and therefore does not wrong future people. In the same-number case however there is no harm to prevent, according to the Counterfactual Condition, despite the fact that the children who would be affected undergo a qualitatively identical ordeal as those in the different people case. As I argued above, however, it is very difficult to see what would justify this difference. We should therefore be sceptical about (i).

To avoid this Kumar could opt for alternative (ii): the two programmes are equally worthwhile, not because they prevent harm, but because they prevent people from being wronged. This does not seem very plausible however. Intuitively, the merits of the same-people programme is the effect it would have on future people's well-being, namely that it would prevent harm. But if the merits of this programme is that it prevents wronging, and wronging is not understood in terms of how it would affect future people's well-being, then it is exceedingly unclear what is meant by "wronging". Kumar says that it requires the "violation of certain legitimate expectations" but as we saw above it is not clear what this implies in same-number cases. In the same-people case (programme *A*) it might seem plausible to say that the affected children can legitimately expect us to do what is best for them, that is, to implement the programme. However, the same reasoning could be used *against* the implementation of programme *B* (the same-number case) since not implementing this programme would be to do what is best for those who would exist if we do not implement this programme. So it is unclear how we on Kumar's notion of wronging can say that the two programmes are equally worthwhile.

Alternatively, Kumar could opt for alternative (iii): the programmes are equally worthwhile because they prevent harm and therefore wrongs. This alternative faces the same difficulties as the rights-approach discussed above. If he were to take on board the view that we can harm people in same-number cases then he would be required to develop an analysis of harm which can account for this judgement. That is, we are led back to defending some version of the Harm Principle against the non-identity problem.

Finally, Kumar could opt for alternative (iv): the programmes are equally worthwhile because one prevents harm (the same-people case) while the other prevents wrongs (the same-number case). However, this alternative suffers from similar problems as (ii). It is not clear that the same-number programme would prevent any wrongs. Whether the same-number programme is implemented or not, it is not clear that the people who will exist because of this choice will have any of their legitimate expectations violated.

This shows that wronging is a poor substitute for harming. Appealing to wronging has no clear advantage over the Harm Principle when it comes to solving the non-identity problem. In short, Kumar's view faces the following dilemma. Either, wronging a person is understood in terms of how an act affects that person's well-being or it is not. If wronging is not understood in terms of well-being then wronging becomes much less plausible as a solution to the non-identity problem. Intuitively, what is objectionable about the young girl's choice has to do with how she affects her child's well-being and the merits of the two medical programmes are how they affect future people's well-being. On the other hand, if wronging is understood in terms of well-being then it does not have any advantage over the Harm Principle as a solution to the non-identity problem. What makes the girl's choice impermissible is that she wrongs her child if she does not wait. This is supposed to be true in virtue of how she affects this child's well-being and not in virtue of what she could have done for some other child. However, this claim faces the same problem as the Harm Principle but in slightly different terms: how can a person be wronged by an act if she would not have been better off had the act not been performed?

An alternative approach is to understand wronging in terms of *desert*. This way of looking at wronging seems not to be vulnerable to the objections just made, and might be what some philosophers who are attracted to the idea have in mind.

There are mainly two ways in which one could take this view. On the first one specifies an amount of well-being and say that people deserve at least that much well-being. One can then say that a person has been wronged if and only if she gets less well-being than she deserves.[26]

A problem for this view is where to place the desert-level. If it is placed high then many future people, even those who are relatively well-off, would be wronged by being brought into existence. This seems counter-intuitive. If it is placed low, however, then the view loses most of its bite since few acts would wrong future persons. Striking this balance is therefore one of the main difficulties for this version of the desert-view.

One way to resolve this issue would be to investigate the grounds for desert. This leads us to the second way of understanding the desert-view. Here the

[26] See Feldman (1997, part III) for the idea that people deserve a certain level of well-being. I will ignore a version of this view where a person has been wronged unless her actual well-being exactly matches her deserved well-being. This view is too implausible.

idea is that people do not deserve some specific amount of well-being but rather that they deserve certain particular goods, and perhaps the absence of certain evils. This version of the desert-view comes closer to the view discussed above, where wronging was understood in terms of rights, and points similar to those which were made against rights apply here as well. We should ask: when is a particular condition an evil? The most plausible answer here, it seems, will be that a condition is an evil, and serves as a ground for desert, just when the condition is also a harm. Or, more cautiously, some harms ground desert, though there might be desert which is not grounded in harm. However, in a case like the young girl it seems clear that harm is the most plausible ground for desert and we would therefore be led back to defending the claim that the girl would harm her child if she does not postpone her pregnancy. This weaker thesis seems to me very plausible and, if correct, shows that sidestepping the non-identity problem by bringing in desert might work in some cases but not all.

In short, the objection against appealing to wronging as a solution to the non-identity problem is that it either (i) neglects the importance of individual well-being, in which case it is implausible as a solution to the non-identity problem or (ii) that it takes well-being into account, in which case the approach has no significant advantage over appealing to harm. One possibility which should be mentioned in this context is to follow Melinda Roberts and define wronging directly in terms of well-being.[27] On her view, to wrong someone is roughly to fail to maximise that person's well-being. However, as Roberts herself notes, the difference between harming and wronging then turns out to be largely a terminological matter.[28] The mere introduction of this distinction will not solve the non-identity problem and the merits of Roberts' account of wronging will have to be evaluated in the same way as one has to evaluate accounts of harm which purport to solve the problem.

### 1.3.3 Harming

The two approaches to the non-identity problem, conditional duties and wronging, both seem to be problematic in one way or another. Though the problems I have highlighted for these views may not be conclusive against them, they show that the person-affecting intuition is not so easily preserved, and the non-identity problem not so easily dispatched, as some philosophers have thought.

To solve the problem in person-affecting terms it therefore seems reasonable to return to the argument against the Harm Principle:

[27] See Roberts (1998, 2003*a*,*b*). Roberts has in her latest works (2010, 2011) dropped the term "wronging" altogether. We will have reason to return to Roberts' view in chapter seven.
[28] See Roberts (2011, p. 337).

(1) An act harms a person if and only if that person is worse off than she would have been had the act not been performed (the Counterfactual Condition).
(2) It is not the case that a person with a life worth living is worse off than she would have been had she not existed at all.
(3) Therefore, a person who would have a life worth living is not harmed by being created.
(4) If a person who would have a life worth living is not harmed by being created, and the Harm Principle is true, then it would be more worthwhile to cure people of a handicap (invest in programme *A*) than to prevent this handicap from occurring (invest in programme *B*).
(5) However, programme *A* and *B* are equally worthwhile (the no-difference view).
(6) Therefore, the Harm Principle is false.

The solutions from the previous sections tried to solve the non-identity problem by rejecting the Harm Principle and replacing it with something else (conditional duties or wronging). The alternative route, which I will explore in this section, is to see if any of (1) to (5) can be questioned.

Most, if not all, of the premises in this argument have been questioned at some point so let me comment on those parts of the argument which I will accept without any further discussion. To some, the most controversial premise in the argument is probably (5). It is important to remember however that (5) is weaker than the more general claim that identity does not make a difference. All it says is that in this particular case, the two medical programmes are equally worthwhile. As I argued above, this claim seems very difficult to deny. I will therefore assume this premise.

We should also grant (4). As I argued above, it is not very plausible that there is some further condition which could explain why the two medical programmes are equally worthwhile if we assume the Counterfactual Condition and the Harm Principle. This leaves us with premises (1) and (2).

First, consider (2). It has been argued that one cannot meaningfully compare the value of existence for a person with the value of non-existence for that person. After all, if the person does not exist then there is no person for which non-existence can have value.[29] However, the argument leaves it open whether such comparisons can meaningfully be made. All that premise (3) commits one to is that *it is not the case that* a person with a life worth living is worse off than she would have been if she had not existed. This claim is compatible with a wide range of views on the value of existence. For example, it is compatible with a life worth living being *better* for the person than non-existence as well as the two states being incomparable. All that it excludes is that a person with a life worth living is *worse* off than she would be had she not existed, and this claim seems very plausible.

[29] See for example Broome (1999, p. 168).

This leaves us with (1). This condition for an act to harm a person has been rather popular but has also met a lot of criticism. In the next chapter I will consider this condition in greater detail and its plausibility. If there are sufficiently strong reasons for rejecting this analysis of harm then the argument against the Harm Principle outlined above would fail.

## 1.4 Summary

In this chapter I have argued that the argument against the Harm Principle, based on the non-identity problem and the no-difference view, rely on a certain analysis of harm: the Counterfactual Condition. I have argued that alternative person-affecting solutions to the non-identity problem, conditional duties and wronging, do not have a clear advantage over the Harm Principle. If anything, these alternative solutions seem to face the same problem as the Harm Principle. I argued that the most plausible way of approaching the problem is therefore to consider the Counterfactual Condition and whether it is a plausible analysis of harm.

However, merely showing that the Counterfactual Condition is not plausible is not sufficient to vindicate the Harm Principle. First, one would also have to develop an alternative analysis of harm which actually solves the non-identity problem. That is, an analysis of harm such that the young girl's choice, or choosing Depletion, actually involves doing harm. Second, the analysis must be compatible with the no-difference view. As I have argued in this chapter an analysis of harm should not make a difference between same-people and same-number cases. Third, the analysis must be intuitively acceptable. We have a rough idea about what is and what is not harm and an analysis of the concept should not stray too far from this pre-theoretical idea.

A brief comment regarding the third condition. As we will see, people, and philosophers no less, tend to have very diverging intuitions when it comes to harm. It is reasonable to expect that an analysis will not be able to satisfy all of them. However, it is possible, or so I will argue in the next chapter, to describe a "basic structure" which incorporates a couple of central features which have a good claim to being essential to the pre-theoretical concept of harm. However, this basic structure is too thin in order to determine any particular analysis. In this thesis I will give priority to the first and the second constraint and my aim is therefore not to give an analysis of harm which fits our pre-theoretical ideas about harm perfectly. Rather, the aim is to find an analysis of harm which solves the non-identity problem and which satisfies the no-difference view while not being too revisionary.

# 2. The Counterfactual Condition

In the previous chapter I argued that in order to defend the Harm Principle against the non-identity problem we should take a closer look at the Counterfactual Condition:

> *The Counterfactual Condition*: An act harms a person if and only if that person is worse off than she would have been had the act not been performed

This claim is the weakest premise in the argument against the Harm Principle. If we can show that it is not plausible then the argument against the Harm Principle based on the non-identity problem and the no-difference view fails. A successful argument against this condition would not establish the Harm Principle, however, since the challenge of accounting for same-number cases, within the constraints of the no-difference view, would still stand. However, the positive argument against the Harm Principle would have been disarmed.

## 2.1 Distinctions

Before discussing the Counterfactual Condition there are a couple of useful clarifications and distinctions to be made. Ordinary harm-talk is extremely diverse and the following is not an attempt to completely map ordinary discourse but only to bring out a couple of distinctions. The Counterfactual Condition is a part of an analysis of the concept "harm" and there are several points to be made regarding analyses of harm in general, and the Counterfactual Condition in particular, which will be important both when evaluating the Counterfactual Condition and when discussing rival views in later chapters.

### 2.1.1 The ontology of harm

To start things off we should distinguish between doing harm and harmfulness. The former, doing harm, is a relational property which can be attributed to wide range of things. For example, smoking (an act) does harm because of how it affects people. The relation here is between the act of smoking and the smoker, and sometimes other people as well. Similarly, an earthquake (an event) does harm because it has certain consequences for those affected by it. Harmfulness, on the other hand, is a non-relational property which is roughly

synonymous with "injury". States of affairs and events can be harmful in this sense without doing harm, that is, without having harmful consequences.

The distinction between these two is often not clear cut and many harmful states or events do harm as well. Furthermore, if an event or state of affairs does harm then this typically means that it has harmful consequences. It is important to note however that there is a conceptual difference between the two and that when analysing harm we have to be clear about which of these two senses we are trying to analyse. If our aim is to analyse doing harm then we should take into account the plausible claim that doing harm is to cause something harmful. Also, we should be careful not to object to an analysis of doing harm that it does not capture our intuitions about which states of affairs or events are harmful. If, however, the aim is to analyse harmfulness then we should consider in virtue of what certain states of affairs and events have the non-relational property of "harmfulness".

A further ontological question concerns which ontological category doing harm should be attributed to. In the Harm Principle, harm is being attributed to acts. In this thesis I will understand acts as a species of events in the sense that the performance of an act, $\phi$, is just the occurrence of an event. An analysis of when an act $\phi$ performed by some person, $a$, harms someone else, $b$, is therefore an instance of the more general analysis of when an event $e$ does harm to $b$.[1]

Which ontological category harmfulness is properly attributed to is not so clear. States of affairs, events and concrete objects are all possible candidates. Here I will make the assumption that states of affairs are the primary ontological category for harmfulness. Attributions of harmfulness to events or concrete objects can, it seems to me, be rendered in terms of states of affairs.[2] This assumption is mainly to simplify the exposition and does not have any significant effect on what I will argue.

With respect to the Harm Principle it should be obvious that it is a principle about the moral relevance of doing harm. For an analysis of harm to be relevant to the Harm Principle it must therefore be an analysis of doing harm rather than harmfulness. The Counterfactual Condition, then, should be understood (and usually is understood) as a condition for when an act (or event) does harm rather than when a state of affairs is harmful. We should also note the following plausible connection between doing harm and harmful states of affairs: an act (or event) does harm only if it has a harmful effect.

---

[1] In ordinary language we also say that *people* do harm. I will in this thesis understand the claim "*a* harmed *b*" as an elliptic way of saying that *a* performed some act which did harm to *b*.

[2] This is not a trivial claim, to be sure. Roughly the idea is that for both objects and events we can construe states of affairs to work in their stead. An object $x$ for example is harmful if and only if the state of affairs "$x$ exists" is harmful. Regarding events, one suggestion is that one can for every occurring event $e$ construe the state of affairs "$e$ occurs".

### 2.1.2 Harm and well-being

With the distinction between harm as a relational property, doing harm, and as a non-relational property, harmfulness, it might be tempting to think that this is a distinction between *extrinsic* and *intrinsic* harm.[3] While doing harm is obviously extrinsic, we should not assume that harmfulness is necessarily an intrinsic property. Poverty, for example, can be quite harmful. But, one of the things which makes poverty harmful is plausibly what it *prevents* a person from achieving, rather than what it is like to be poor. Poverty can of course be intrinsically harmful but it can also, I suggest, be extrinsically harmful. Other examples of extrinsic harmfulness are cases of deprivation. When a person is deprived of some good, or the ability to achieve that good, then we might say that that person is in a harmful state. However, this state does not seem to be harmful because of any of its intrinsic features.

We should therefore not say that the distinction between doing harm and harmfulness is just the distinction between extrinsic and intrinsic harm because harmfulness can be an extrinsic property. However there is a distinction to made here. In value theory a distinction is sometimes made between final and instrumental value. Something has instrumental value if and only if it is valuable because it has consequences which are finally valuable. To have final value, however, is to be valuable as an end and not merely as a means to something else which is valuable.[4] A way to distinguish between doing harm and harmfulness would be to use this distinction between final and instrumental rather than intrinsic and extrinsic. With the distinction between final and instrumental harm we can say that doing harm is harm in the instrumental sense while harmfulness is harm in the final sense. To say that $x$ is an instrumental harm is to say that it leads to something which is a final harm. A final harm, on the other hand, is not harmful because of its consequences but because of its nature.[5]

This way of characterising the difference between final and instrumental harmfulness leaves it open what, exactly, final harms are. Traditionally, final harms have been thought of in terms of "interests".[6] Roughly, the traditional analysis of final harms hold that they are "setbacks" to a person's interests while instrumental harms are the causes of such setbacks. The aim of this terminology is to capture one central aspect of final harms: final harms matter to the one who suffers them in the sense that they are *bad for* the person who suffers them. This connection with "prudential value", value-for, suggests that we can think of final harms as negative components of a person's well-being. Well-being, as I will use the term, is what makes life good (or bad) for the

[3] Bradley (2012) suggests this terminology to separate the two.

[4] Regarding the difference between final and intrinsic value, see Korsgaard (1983) and Rabinowicz & Rønnow-Rasmussen (2000).

[5] "Nature" should of course not be understood as being limited to a thing's intrinsic properties.

[6] See for example Bayles (1976) and Feinberg (1986, 1987).

person whose life it is.[7] We do not need to take a stand here on what exactly well-being is, whether it is pleasure, preference satisfaction or something else. Here I am more concerned with the formal properties of harm and to a lesser extent the substantial ones about what it is that makes life go better or worse. The traditional view of final harms as setbacks to interests assumes one particular view of well-being but an analysis of harm should as far as possible be compatible with different views on what well-being is.[8]

In short, instrumental harms are harms in virtue of their consequences while final harms gain their status in virtue of their nature. When we say of some act that it is harmful we typically mean that it is harmful because of its consequences. That is, acts are usually not harmful in the final sense but only in the instrumental sense. The qualifier "usually" is needed here however. Some acts, such as extremely offensive or insulting acts could be considered harmful in the final sense. Whether such examples should be classified as final harms is a question which I will not attempt to answer in this thesis, it will have to be settled by a theory of well-being. We can however note that the following claim seems plausible in the light of the connection between harmfulness and well-being: a state of affairs is harmful in the final sense if and only if it is bad for someone in the final sense.

The Counterfactual Condition is of course compatible with the distinction final-instrumental harm, but it should be emphasised that the Counterfactual Condition applies only to doing harm and not harmfulness.[9] Only applying the condition to doing harm and not harmfulness, is by no means a weakness but rather a strength. It enables a defender of the Counterfactual Condition to say, for example, that while the young girl's choice not to postpone her pregnancy does not harm her child, it is still *bad for* the child to have a bad start in life.

---

[7] It is important to distinguish well-being, or prudential value, from other forms of value such as aesthetic or ethical value. To say that something promotes a person's well-being is not necessarily to say that it is good ethically, or aesthetically. Similarly, a life which is good for me, a life where I have high well-being, might be an aesthetically bad life. See also Sumner (1996, pp. 20–6).

[8] It would be appropriate to say something about what I take a plausible theory of well-being to be. There are mainly two constraints which, on my view, a theory of well-being will have to satisfy in order to qualify as initially plausible. First, the theory should imply that it is possible to be mistaken about one's own well-being. Though the first-person perspective might give one a privileged position epistemically, a person is not infallible in her judgements about her own well-being. Second, what contributes to a particular person's well-being is, to some extent, person-relative. That is, it is possible that a particular thing is good for me but not for you. This constraint does of course not rule out that there are objective facts about well-being but it does, for example, rule out a view which identifies well-being with intrinsic impersonal value (this view is briefly discussed in chapter three). These two constraints only rule out very crude views about well-being and should not be very controversial. Filling in the details is however not a task for this thesis.

[9] Insofar as the distinction is made, philosophers who discuss the Counterfactual Condition intend it to apply only to instrumental harm. See Feinberg (1986, pp. 148–9) and Bradley (2012).

Likewise, one can say in the case of the Specks that no one *harmed* their child but that it is bad for the child to be born with neurofibromatosis.[10]

### 2.1.3 Partial and total harm

When discussing harm it is common to distinguish between a *partial* or *pro-tanto* sense and a *total* or *all things considered* sense. In the partial sense harm is used to convey that an act or decision has consequences which are to some extent harmful while the total sense is used when we say that an act or decision has consequences which are on the whole harmful.

To illustrate this distinction, consider the following case:

> *Surgery*. Black saves White's life by amputating White's leg. White suffers intense pain but had Black not amputated the leg then White would have died.

Removing a limb, even by surgical means, is something which under normal circumstances would be considered a paradigmatic example of doing harm. But, in this case circumstances are not normal because the amputation is necessary in the circumstances for saving White's life. One way to view this situation is to say that while amputating harms White it is only a *partial* harm. It is something which, taken by itself, harms White. However, because amputating saves White's life it is not to do harm in the *total* sense.

The relation between these two senses of harm seems to be fairly straightforward: harming in the total sense is a function of harming and benefiting in the partial sense. In the case of amputation, for example, we weigh the partial harm of loosing a limb against the partial benefits (if there are any) of loosing a limb. If the partial harms outweigh the partial benefits, then we say that the amputation harmed the person in the total sense. Otherwise, it only harms her in the partial sense.

The Counterfactual Condition tends to be viewed as a condition for total rather than partial harm.[11] In Surgery, for example, philosophers who have discussed the Counterfactual Condition take it to imply that amputating does not harm White because not amputating would be all-things-considered worse for White. They would not deny, or would not have to deny at least, that amputating harms White in some sense but they would claim that this is not the sense of harm which the Counterfactual Condition is supposed to apply to.

#### 2.1.3.1 The morally relevant sense of harm

The distinction between partial and total harm invites the question whether it is the total or the partial sense which we should be analysing. Defenders of the Counterfactual Condition usually claim that it is the total and not the partial

[10] If this seems puzzling, see below regarding the morally relevant sense of harm.

[11] See Bayles (1976, p. 293), Feinberg (1986, p. 147), Norcross (2005, p. 150) and Bradley (2012).

sense which is the *morally relevant* sense of harm, and we need therefore to take a closer look at what this claim might mean and whether it is plausible.

One thing one could mean by "morally relevant sense of harm" is that some harm-attributions do not have the right kind of subject. We might want to draw a distinction between harming persons and harming other things, such as the environment, works of art etc. Because well-being is central to the concept of harm it might be argued that things which cannot properly be said to have well-being might yet be harmed in an extended sense.[12] The idea would be that this sense of harm, harm to non-persons, is not morally relevant in the sense that we do not have to take such harms into account in moral deliberation. I mention this possibility mostly to put it aside. It is obvious that even if there is a morally relevant difference between harming, say, the environment and harming a person, this difference is independent of the total-partial distinction.

A better approach is to claim that to draw the distinction between morally relevant and irrelevant harms we need to consider how harm relates to normative principles. Here a natural suggestion is that to analyse harm in the morally relevant sense is to analyse harm as it occurs in normative contexts like the Harm Principle. The claim that it is harm in the total sense which is morally relevant then amounts to the claim that we should understand normative principles like the Harm Principle as referring to harm in the total sense.

It should be emphasised how important the claim that the Counterfactual Condition captures the morally relevant sense is for the argument against the Harm Principle. To see this, consider the following *reductio* of the Counterfactual Condition:

> (1) If the Counterfactual Condition is true, then it is not the case that there is harm done in the case of the young girl.
> (2) There is harm done in the case of the young girl.
> (3) Therefore, the Counterfactual Condition is false.

Premise (1) should not be controversial, and (2) is simply the common sense verdict when faced with same-number cases. The way we ordinarily use "harm" does not stop us from saying that the girl would harm her child if she does not wait. One way to argue against the Counterfactual Condition is then to use the fact that it conflicts with ordinary harm-attributions as a reason for rejecting it as part of an analysis of harm.[13]

The obvious reply from the defender of the Counterfactual Condition is to appeal to the distinction between harm in a morally relevant sense and harm

[12] See Feinberg (1980) for a discussion of what kinds of things can be harmed. Even though Feinberg frames his discussion in terms of "interests" his discussion applies to harm as well.

[13] Harman (2004) seems to suggest an argument along these lines: "I claim that *causing* pain, early death, bodily damage, and deformation is harming. We do not need a complete analysis of what it is to harm, in order to reach this conclusion; we can hold that these are clear cases of harm" (Harman 2004, p. 92).

in a "wider" sense: in (1) harm is used in the morally relevant sense while in (2) it is used in the wider sense.

How does one substantiate the claim that it is the total sense rather than the partial which is morally relevant? Sumner (1996) has discussed this problem for a related concept: welfare. Sumner's view is that when analysing a concept like welfare we strive for two things: descriptive adequacy and normative adequacy. The first of these is achieved if the analysis fits ordinary discourse, the way "welfare" is used in ordinary language. The second aim, that of normative adequacy, is achieved if the analysis makes a specific normative principle which refers to welfare plausible. Sumner's view is that we should give some priority here to descriptive adequacy and especially to the "pre-analytic core" of a concept. What this means is that there are certain parts of ordinary discourse, the core, which it is more important for an analysis of welfare to be faithful to than other areas. However, Sumner grants that there might be several competing analyses which satisfy descriptive adequacy equally well. Here is where normative adequacy enters: "when the evidence provided by our ordinary experience is indeterminate or inconsistent, then there is a time for shaping a theory of welfare to fit some favoured normative niche" (Sumner 1996, p. 19).

The idea that there is a pre-analytic core to a concept which should play an important role when analysing a concept can be compared with what Smith (1994) refers to as the "platitudes" regarding a concept. Smith's view is that when analysing a concept there are a number of propositions involving the concept which have a "*prima facie a priori* status". According to Smith, these platitudes are central to the concept in two ways. First, the platitudes are central to mastering the concept. A person who masters a concept treats these propositions as true; she is at the very least disposed to accept them as true. Second, an analysis "should give us knowledge of all of the relevant platitudes [...] that is, the maximal consistent set of platitudes constitutive of mastery of the term" (Smith 1994, p. 31).

Sumner and Smith share the idea that when analysing a concept there are certain propositions which we should pay special attention to. What is characteristic of these propositions is that they are constitutive of the meaning of the concept: a person who uses the same *term* but in a way which is inconsistent with these propositions is talking about something else than a person who uses the term in accordance with them.[14] In what follows I will use Sumner's terminology of "pre-analytic core" but this is merely a matter of convenience.

If we apply this to harm we can say that to substantiate the claim that it is the total sense of harm which is morally relevant one would have to show that the total sense is central to the pre-analytic harm-discourse, especially that part of the discourse which pertains to The Harm Principle or similar normative principles. Alternatively, one could argue that the pre-analytic core

[14] Jackson (1991, pp. 31–42) emphasises the role platitudes have to fix the discourse.

and the relevant part of the harm-discourse is indeterminate with respect to the total-partial distinction. In that case, we can choose whichever best fits the normative role we have in mind for harm. We should therefore try to formulate, as precisely as we can, what the pre-analytic core of "doing harm" is.

I have already mentioned two candidates for membership in the pre-analytic core: an act (or event) does harm only if it has a harmful effect and a state of affairs is harmful in the final sense only if it is bad for someone in the final sense. These two putative platitudes about doing harm can be captured somewhat more stringently by what I will call the "basic structure" of harm:

> *a* harms *b* only if
> (1) *a* performs an act, $\phi$,
> (2) *b* is in a state *S* which is bad for *b* in the final sense.
> (3) $\phi$ is responsible for *S*'s obtaining.

These three claims seem very plausible. The first condition should be trivial, provided that we confine the analysis to acts. This condition serves to distinguish an analysis of when an agent harms someone from a more general analysis of when natural events do harm as well as distinguishing mere bodily movements from actions proper. The second condition captures the idea that doing harm requires a certain kind of effect as was spelled out above. The third condition seems necessary since the harmful state must in some relevant sense be *attributable* to *a*'s $\phi$-ing. The mere fact that there is a harmful effect and that someone performs an act is of course not enough for the act to do harm; the act must be related in the right way to the harmful effect in order for the act to do harm.[15]

It should be emphasised that I intend these conditions to leave a lot of room for further analysis. For example, regarding (1) it is a further issue whether omissions are acts in the relevant sense as well as what distinguishes acts from mere bodily movements. In (2) there is no need to settle the exact boundaries of which states are bad for *b* in the final sense, that should be left to a theory of well-being, and in (3) we can leave it open exactly what it takes for an act to be responsible for a certain effect in the intended sense. For a complete analysis we would have to settle these issues but in order to capture any platitudes about doing harm we should not beg any questions regarding these notions. I will therefore, for the moment, leave it open how (1)-(3) are to be understood more precisely.

In addition to the basic structure we could also add a list of paradigmatic instances of doing harm such as killing, torture and so on. However, all such

[15] Why not simply say that $\phi$ causes *S* to obtain? The reason is that it seems plausible that the relevant relation is some kind of dependence relation (*S*'s obtaining depends on $\phi$'s occurring), but is not clear whether all dependence relations are causal. In the light of this it seems plausible to hold that an act can be responsible for things which it does not cause. I will discuss the causal view of responsibility in chapter five.

paradigmatic examples would need a *ceteris paribus* clause attached to them because, as we saw in Surgery, what would be a clear case of doing harm in one context is not a clear case in other contexts. Even death, which might seem like a paradigmatic example if ever there was one is, by many philosophers at least, not taken to be a harm (or a "misfortune") whenever it occurs.[16] It is also difficult to find other features which could be taken as a part of the core. Candidates such as that harming involves certain intentions or that it is a matter of violating rights are such that we could reasonable disagree about whether they are necessary for doing harm. In the case of intention, for example, it is not clear whether we are simply confusing the relevance of intention to harm with the moral importance of intending harm.[17]

This leaves room for normative considerations to play an influential role when analysing harm. As a rule of thumb, I suggest that we can depart from descriptive adequacy only if it can be motivated independently of achieving greater normative adequacy. Simply trading descriptive for normative adequacy threatens the analysis to be about something else rather than about harm.

Let us now return to the claim that it is the total sense of harm which is the morally relevant sense. It should be clear, I hope, that this claim cannot be decided either way by merely looking at the pre-analytic core. Both the partial and the total sense are compatible with there being a close connection between harm and well-being. A case like Surgery suggests that it is the total sense which is the morally relevant one but we also use harm in the partial sense. For example, it would not be unreasonable to say that Black harms White in the Surgery case if one adds that there are compensating benefits associated with inflicting that harm. Whether it should be accepted depends to a large extent on what normative role one has in mind.[18] For our present purposes, it should be noted that the Harm Principle could be understood in either partial or total terms. Reading it as referring to total, or partial, harm does not make it absurd on its face. We should therefore accept that the tools at our disposal, the pre-analytic core and the Harm Principle, underdetermine whether it is the total or the partial sense which is the morally relevant sense of harm.

[16] See for example Feldman (1991), Nagel (1991), Feit (2002) and Bradley (2009).

[17] Bradley (2012) argues that separating intuitions about what one ought to do from intuitions about harm is important if we are to avoid "moralistic fallacies" of this kind.

[18] Kagan (1998, pp. 86–8) suggests a similar approach to the question which sense of harm is morally relevant. He distinguishes between "global" and "local" harm and indicates that both can be used to indicate harm in a morally relevant sense.

## 2.2 Objections to the Counterfactual Condition

The basic intuition behind the Counterfactual Condition is that harm should be analysed in comparative terms. The intuition is that, as Parfit puts it, "[i]f what we are doing will not be worse for some other person, [...] we are not, in a morally relevant sense, harming this person" (Parfit 1984, p. 374). This seems quite plausible. We do not only think, as I suggested above, that to do harm is to make some person suffer a state of affairs which is bad for her. It also seems plausible that to do harm is to make a person's life go *worse*. Doing harm, it seems, is not only a matter of making people badly off, in some respect. Doing harm also has a "contributive character" in that it is to make someone's life go worse.[19]

This intuition is however ambiguous because we have not specified what the relevant comparison is regarding "worse for". Worse than what? According to the Counterfactual Condition, an act harms a person if and only if the person is worse off than she would have been had the act not been performed but there are many comparisons which could be made, and hence many analyses of harm which satisfy the intuition but which give drastically different results.[20] Something needs to be said in favour of adopting a counterfactual comparison and not some other.

One suggestion which satisfies the underlying intuition is the Temporal View. On this view to harm a person is to make that person worse off than she was before the harm came about. More precisely:

> *The Temporal View*: An act $\phi$ harms a person $a$ if and only if $\phi$ causes $a$ to be worse off at time $t$ than she was before $t$.[21]

There are mainly two counterexamples to this view. The first is how the Temporal View is to account for cases where an act makes a person worse off than she was, but better off than she would have been had the act not been performed.[22] Consider a variation of the amputation case. If Black does not amputate White's leg then White will die while if Black amputates then White will live, but will be slightly worse off than she was prior to the amputation.[23]

---

[19] The connection, if there is any, between the badness of a state of affairs for a person and the difference, the contribution, that state of affairs makes to a person's life, will be further explored in chapter four.

[20] The Counterfactual Condition is sometimes formulated in terms of "could" rather than "would". This difference is significant when there are more than two alternatives for an agent to choose from. In what follows I will only consider simple cases where there are only two alternatives and formulating the Counterfactual Condition in terms of "could" rather than "would" will therefore not have any effect on what I will argue in this chapter.

[21] This view has many critics but few defenders. See however Perry (2003).

[22] This argument is advanced by for example Norcross (2005).

[23] If this seems unlikely, we can assume that White has some kind of infection in her leg which she does not suffer from now, but which will in the future result in her death.

The temporal condition implies that Black would harm White by amputating, but this seems counter-intuitive.

In defence of the Temporal View it could be argued that we should read it as a condition for partial harm rather than total harm. Perhaps we should say that an act harms someone, partially, only if the act makes that person worse off in some respect than she was. This will not do however. In order for this reply to work we need to identify some respect in the amputation-case in which amputating makes the person better off than she was. Otherwise, the Temporal View still seems to imply that amputating would harm the person all things considered. The most straightforward way would be to say that it is the fact that amputation saves the person's life which accounts for this but it is unclear whether a defender of the Temporal View can say this. When saving someone's life it seems very strained to say that we are making that person better off than she was.

The second counterexample to the Temporal View is when a person is prevented from receiving a benefit. As we have just seen, the Temporal View seems to have difficulties with accounting for putative harms which involve a prevention of something good. Many such cases are however clear examples of harm. For example, if Black conspires in order to prevent White from getting a prize which White would otherwise have been qualified for then it seems clear that Black harms White (supposing that it would be good for White were she to get the prize of course).[24]

A way for the Temporal View to reply to the objection could be to appeal to a view where it is worse for a person to have a smaller chance of receiving a benefit. It could then be claimed that preventing a person from receiving a benefit is to lessen that person's chances of receiving that particular benefit. Preventions of benefits would then count as doing harm since they lessen a person's chances at receiving a particular benefit.[25] To illustrate, consider a child who is prevented from receiving a normal education. As this child gets older it gets harder and harder for her to learn how to read, for example. In this case it seems plausible to say that preventing the child from getting a normal education lessens the child's chances of ever learning how to read since it is easier to learn such things when you are young. On the reply under consideration, having a smaller chance of learning how to read is worse for the child and preventing her from receiving a normal education makes her worse off in the temporal sense.

While it seems plausible that there are cases like the one just described it is not the case that all preventions of benefits can be dealt with in this fashion. Some preventions of benefits leave a person's chances of receiving the benefit constant, or may perhaps even increase them, if the benefit in ques-

[24] For further examples of this kind, see Feinberg (1986, p. 149), Holtug (2002, p. 368) and Hanser (2008, p. 429).

[25] See Thomson (2011, pp. 444–5). It should be noted that Thomson rejects the Temporal View but for other reasons than ones having to do with failures to benefit.

tion is not something which, as with learning to read, gets more difficult to receive over time. Consider Black's conspiring against White again. Suppose that the prize is "employee of the month" and that it is a policy to award it to a different employee every month. Then it is not the case that by preventing White from becoming employee of the month *this month* Black thereby lessens White's chances of becoming employee of the month. Quite the contrary, White's chances of getting the prize over the next months will be greater since the number of available candidates will become smaller and smaller.

The fact that preventions of benefits are not even putative harms on the Temporal View gives us decisive reasons to reject it. Many such examples are clear instances of harm and the Counterfactual Condition looks more plausible in this respect. Indeed, not only preventions of benefits but also failures to benefit are harms on this view. Whether this makes the view too wide is something I will now turn to.

### 2.2.1 Failures to benefit

A common objection against the Counterfactual Condition is that it makes mere failures to benefit into harms.[26] Suppose I am asked to donate a kidney to my neighbour. If I refuse she will live a decent life while if I comply with the neighbour's request then she would be much better off. In this case it seems absurd to say that I would harm my neighbour if I refuse. Compare this with a case where I steal a kidney from a perfectly healthy person making her worse off than she would have been had I not stolen the kidney. The result in both cases is that someone ends up with one kidney less than she would otherwise have had. The Counterfactual Condition is therefore fulfilled. But there seems to be an important difference between the two. In the first case I do not harm anyone while the second seems to be a paradigmatic example of doing harm. There are however a number of things which can be said in defence of the Counterfactual Condition here.

We have already seen that with respect to certain preventions of benefits, such as preventing a person to develop her full potential, it is not counter-intuitive to say that the person has been harmed by being prevented from receiving the benefit. In defence of the Counterfactual Condition it could then be claimed that in cases like the one where I do not donate a kidney to my neighbour our intuitions are not very reliable because there are other features of these cases which make our intuitions unreliable. The defender of the Counterfactual Condition could therefore claim that we should depart from descriptive adequacy in the kidney case because our judgement that not donating a kidney is not to do harm is not a reliable one.

If we compare the case where I do not donate a kidney with the case where I prevent a person from developing her full potential, one difference is that

[26] See for example Shiffrin (1999, p. 121), Harman (2004, p. 98), Hanser (2008, p. 428) and Bradley (2012).

in the first case I refrain from doing something while in the latter I actively prevent something which would have occurred in the normal course of events. Our tendency to make different judgements about these cases can perhaps be explained by our tendency to think that there is a morally relevant difference between acts and omissions, even if we would reject this view were we to consider it in a more critical fashion. In defence of the Counterfactual Condition it could then be claimed that our intuitions in cases of failings to benefit are usually influenced by a tendency to make a distinction between acts and omission, but that this distinction is not a relevant one.[27]

Even if one grants the claim that the distinction is irrelevant to an analysis of harm, and that our intuitions are influenced by a tendency to affirm the distinction, this argument seems questionable. If our intuitions about harm and harming are influenced in this way one would expect that our judgement regarding the kidney case would change when we come to realise this. However, it is not so clear that the counter-intuitiveness of saying that I harm my neighbour when I fail to donate a kidney disappears when I realise that this is just because it is a case involving an omission. A more plausible diagnosis of the difference in our intuitions here is that there is some other feature of the two cases which accounts for our tendency to make different judgements about them which may be relevant to an analysis of harm.

Furthermore, it is unlikely that this strategy would work for all cases where the Counterfactual Condition implies that a person does harm by failing to benefit someone. In the kidney-example there is something to be said in favour of the claim that I would harm my neighbour by not donating a kidney because it would be to deprive my neighbour of a benefit, and depriving a person of a benefit is at least sometimes to harm that person. But, the Counterfactual Condition implies that I would harm my neighbour even in cases where I fail to do what is best for my neighbour. For example, suppose that my neighbour would be better off if I donate a kidney, but that she would be even better off if I donate a lung, then the Counterfactual Condition implies that I would harm my neighbour by donating a kidney. But this seems absurd. Perhaps I *should* donate a lung, but it seems very counter-intuitive to say that I would harm my neighbour by merely parting with a kidney.

As this example shows, the Counterfactual Condition clashes with the second condition in the basic structure; that an act does harm only if it has a harmful effect. The Counterfactual Condition also clashes with the condition that an act does harm only if the act is responsible for a harmful effect. According to the Counterfactual Condition, all that is relevant is that a certain counterfactual claim is true in order for an act to do harm. According to the Counterfactual Condition, if an act makes a person worse off than she would otherwise have been then the act does harm, even if this is because of a mere coincidence.[28] The Counterfactual Condition is therefore very far removed

[27] See Kagan (1989, ch. 3) who argues against the relevance of the distinction.

[28] I discuss this problem for the Counterfactual Condition at greater length below.

from the ordinary concept of harm and does not do very well with respect to descriptive adequacy. However, as was noted above, departures from descriptive adequacy are sometimes acceptable, especially when such departures can be independently motivated. Paying attention to the distinction between doing and allowing can, in some cases, motivate these departures from descriptive adequacy but it will not work in all cases.

Could it be argued that the Counterfactual Condition should be accepted, despite its counter-intuitive implications, because of its normative adequacy? First, as I argued above we should be careful with simply trading descriptive for normative adequacy. The pre-analytic core of a concept especially should be respected as far as possible. Second, as the non-identity problem and the no-difference view shows, it is clear that the Counterfactual Condition is not normatively adequate because it fails to make normative principle which refer to harm, such as the Harm Principle, plausible. We should therefore conclude that the Counterfactual Condition is not a plausible analysis of harm as it occurs in the Harm Principle.

### 2.2.2 Reformulating the Counterfactual Condition

These objections against the Counterfactual Condition suggest that it is too strong. There is, however, a weaker version of the condition which might be more plausible. As I will argue, this weaker version of the Counterfactual Condition avoids the objections raised above while still capturing the plausible claim that to harm someone is to make that person's life go worse.

According to the weaker version, the Counterfactual Condition is only a necessary condition for doing harm:

> *The Weak Counterfactual Condition*: an act harms a person *only if* that person is worse off than she would have been had the act not been performed.

The Weak Counterfactual Condition avoids the objections raised at the end of the previous section because it is only a necessary condition. It is, for example, compatible with the basic structure.

The weak version is what at least some proponents of the stronger version seem to have in mind. For example, in Feinberg's analysis of when *a* harms *b* he includes the Weak Counterfactual Condition but apart from this it also includes as a necessary condition that "[*a*]'s action is the cause of an adverse effect on [*b*]'s self-interest (a "state of harm")" (Feinberg 1986, p.148). Regarding the distinction between a "state of harm" and doing harm Feinberg writes that

> there is a sense in which "state of harm" is the more fundamental concept, since there can be no act of harming unless a state of harm is its product, whereas we *can* have a state of harm without there being any prior act of harming as its cause. (Feinberg 1986, p. 148).

In the analysis that follows Feinberg makes it clear that the Weak Counterfactual Condition is not a condition for when a particular state is a state of harm but when an act performed by an agent harms someone. What I take Feinberg to be referring to here is the distinction between doing harm and harmful states (or events). Feinberg's distinction between "states of harm" and "harming" is the same distinction which was made above between doing harm and harmful states of affairs (or events).

This also sheds some light on the role of the Counterfactual Condition in an analysis of harm. It is not, as some seem to have thought,[29] a condition for when a state of affairs is a "state of harm". As was noted when I introduced the basic structure of an analysis of harm, a complete analysis should include a condition to the effect that an act does harm to a person only if the act has a harmful effect. However, the Counterfactual Condition as only a necessary condition for when an act does harm is compatible with this part of the basic structure and, furthermore, does not imply anything about how such harmful effects should be analysed. Furthermore, I argued above that a plausible necessary condition for when an act does harm is that the harmful effect must be "attributable" to the act. The Weak Counterfactual Condition is distinct from this condition as well.

A benefit of weakening the Counterfactual Condition is that it us to distinguishing the Counterfactual Condition from questions about responsibility. Because the Weak Counterfactual Condition is compatible with any view about when a harmful effect is attributable to an act it is possible to hold that some acts are responsible for harmful effects without doing harm. In the Surgery-case for example, one could say that performing the amputation has a harmful effect but, because the Weak Counterfactual Condition is not satisfied, amputating does no harm.

Note also that weakening the Counterfactual Condition is that it does not have any relevant impact on the non-identity problem or the no-difference view. The Weak Counterfactual Condition is not satisfied in the cases considered in chapter one (the young girl and Depletion) and it is therefore sufficient to rule out that there is any harm done in these cases. Regarding the no-difference view and the case of the Two Medical Programmes the condition rules out that there can be any harm done by cancelling programme *B* (the same-number programme). It allows, but does not entail, that there is harm done by cancelling programme *A*. It seems plausible however that any further conditions would also be satisfied. There is, for example, an intuitively recognisable harmful state of affairs (the handicap the children would have if the programme is cancelled) and our choice seems, *prima facie*, to be responsible for whether this effect comes about.

[29] For example, Hanser (2008, p. 423) attributes to Feinberg the view that the Counterfactual Condition is a condition for when a state is a "state of harm". This seems to be a mistake since Feinberg clearly distinguishes between "doing harm" and "harmed conditions" or "states of harm".

It therefore seems fruitful to see the Weak Counterfactual Condition as an addition to the basic structure. Placing the Weak Counterfactual Condition in relation to the basic structure we can say that if we suppose that the basic structure is necessary, what is missing for conditions (1)-(3) to be jointly sufficient according to the proponent of the Weak Counterfactual Condition is that $\phi$-ing makes $b$ worse off than she would otherwise have been.

This more sophisticated view seems to have the resources to respond to some of the objections from failures to benefit. In cases where a person would have been better off were an act to be performed, but there is no harmful effect, one would not be harming this person by failing to perform the act. For example, not further increasing the well-being of a person who already is very well off would not be an instance of harming (depending, of course, on how "harmful effect" is spelled out).

The Counterfactual Condition as only a necessary condition is therefore more promising than as necessary and sufficient. However, the weaker version encounters problems of its own which, I will argue, give us reason to reject it and look for a more plausible analysis of harm.[30]

### 2.2.3 Irrelevant consequences

The Weak Counterfactual Condition can be accused of including considerations which are irrelevant to whether an act harms someone. Woodward raises this objection with the following example:[31]

> Victor Frankl seems to suggests that, as a result of his imprisonment in a Nazi concentration camp, he developed certain resources of character, insights into the human condition, and capacities for appreciation that he would not otherwise have had. Let us suppose, not implausibly, that Frankl's mistreatment by the Nazi's was a necessary condition for the richness of his later life, and that, had the Nazis behaved differently toward him, his life would have been, on balance, less full and good. [...] It is Frankl, and not the Nazis, to whom credit and responsibility for his later life are due. (Woodward 1986, p. 809).

In this example the Weak Counterfactual Condition is not fulfilled. Frankl is not worse off than he would have been had the Nazis acted differently, and the obviously counter-intuitive conclusion is that the Nazis did not harm Frankl. It also seems inappropriate, as Woodward points out, to take the benefits of Frankl's later life into account when determining whether the Nazis harmed Frankl.

The Weak Counterfactual Condition is, as I mentioned above, a condition for total harm and one way to defend the condition could be to rely on the distinction between partial and total harm. It seems plausible to claim that with

[30] These objections are, of course, also objections to the stronger version of the condition.
[31] A similar objection is raised by Hanser (1990, p. 60).

respect to some aspects, Frankl is worse off than he would otherwise have been, though he on the whole is not worse off. In Woodward's example, the conclusion drawn from the Weak Counterfactual Condition is that the Nazis did not harm Frankl in the total sense. But, on behalf of the Weak Counterfactual Condition it could be replied that this does of course not rule out that the they harmed him in a partial sense. Furthermore, a defender of the Weak Counterfactual Condition can point to a tendency not to distinguish between partial and total harm in ordinary language as explaining why it might seem counter-intuitive to say that the Nazis did not harm Frankl.[32]

This defence seems to be a small comfort however since it is still counter-intuitive to say that the Nazis did not harm Frankl all things considered. A point which could be raised here in defence of the Weak Counterfactual Condition is that our intuitions in the Frankl-case are influenced by our tendency to condemn the Nazis actions from a moral point of view. However, such moral condemnation should be distinguished from the claim that in the morally relevant sense the Nazis did not harm Frankl. It could then be argued that what the Nazis did was wrong, even though for other reasons than that they harmed Frankl in the total sense.

A more troubling aspect of Woodward's objection for the Weak Counterfactual Condition is that it seems to take irrelevant considerations into account. The benefits of Frankl's later life is not something we would consider as relevant when evaluating whether the Nazis harmed Frankl. Of course, a defender of the Weak Counterfactual Condition can point to cases where we have the opposite intuition.[33] Suppose that Frankl suffers from post-traumatic stress because of what the Nazis did to him. This is certainly something which we are inclined to blame the Nazis for but it is not clear what the relevant difference between this case and Woodward's is. This reply would however be to change the subject. Whether the Nazis are responsible (in the relevant sense) for what happens to Frankl in his later years is one thing, but if the benefits he would enjoy in his later life are such that he would not have enjoyed them had the Nazis done otherwise then they are taken into account by the Weak Counterfactual Condition regardless of what we say about the Nazis responsibility for these benefits (and likewise for harms).

We can make this point in a more formal way. Suppose two events, $e_1$ and $e_2$, would not have occurred had an act, $\phi$, not been performed. Suppose fur-

[32] Could the Weak Counterfactual Condition be modified further so that it is a condition for partial harm? One way to do this would be to apply the condition to *parts* of a life rather than a whole life. On this view, the Nazis harmed Frankl because Frankl was worse off during a period of time than he would otherwise have been during that period. However, this would be a poor analysis of partial harm because several partial harms and benefits can obtain at the same time. For example, an analysis of partial harm should, it seems, entail that there is harm done in the Surgery-case but it is doubtful that this would follow from an analysis of partial harm in counterfactual terms.

[33] See Smilansky (2007, ch. 1) for a nice illustration of how our intuitions tend to go both ways in cases like the ones discussed here.

ther that $\phi$ is only responsible in the relevant sense for $e_1$. Now, whether $\phi$ harms someone, according to the Counterfactual Condition, depends on the impact of *both* events on a person's well-being. But, whether $\phi$ does harm or not would then depend on effects which it is not responsible for. This contradicts the plausible claim that whether an act does harm depends only on the effects of this act. If the appropriate link between an act and a harmful state of affairs is missing, i.e., if the act is not responsible for the state's obtaining, then the act cannot do harm in virtue of the state of affairs' obtaining.[34]

In defence of the Weak Counterfactual Condition it could be claimed that it is a mistake to assume that counterfactual dependence and responsibility do not go hand in hand. To assume that the events above would not have occurred had the act not been performed is just to assume that the act is responsible for those events. While this reply makes the act responsible for both effects, it involves a considerable broadening of the ordinary notion of responsibility. What this reply amounts to is that, in Woodward's example, the Nazis are responsible for the benefits Frankl received in his later years.

The best reply, it seems, for a defender of the Weak Counterfactual Condition here is to be revisionistic and argue that we should accept this broadening of our ordinary notion of harm and responsibility. If we do, then it seems we could also argue against Woodward's example on the grounds that it relies on a mistaken view about responsibility. Understanding responsibility as counterfactual dependence does capture something central to responsibility, namely that we are responsible for *the difference* our acts make. I will not attempt to answer the question whether we should accept this broadening of our ordinary notion of responsibility, but we will have reason to revisit this topic in chapter five. For now I will merely note that the best option for a defender of the Weak Counterfactual Condition seems to be the revisionistic path.

Summing up, while Woodward's example does not damn the Weak Counterfactual Condition it shows that there are certain costs attached. First, it would require us to give up the intuitive judgement that the Nazis harmed Frankl all things considered. Second, the Weak Counterfactual Condition pushes one to adopt a revisionistic account of responsibility. If an event would not have occurred had a certain act not been performed then the Weak Counterfactual Condition takes this into account when determining harm, even if we intuitively would not say that the act is responsible for the effect.

[34] A problem for this restriction, it has been argued, concerns collectives of acts. The restriction seems to rule out that mere participation in a collective act can be to do harm even though one's individual contribution does *not* harm. See for example Eggleston (2000). Similarly, Parfit (1984, p. 70) argues that we should not commit the mistake of saying that if an act is right (wrong) because of its effects, then "the only relevant effects are the effects of this particular act". However, Petersson (2004) has argued, based on Lewis (2000), that there is no need to appeal to mere participation in collective cases and that the restriction should be accepted. I will here assume that mere participation does not matter, but that the notion of an act's effects should be broadened. I discuss this matter further below and in chapter five.

### 2.2.4 Overdetermination and pre-emption

The most common, and most serious, objections to the Weak Counterfactual Condition are cases where the effects of an act are *overdetermined* or *pre-emptied* by another act. Consider the following cases:

> *Pre-emption*. Black poisons White. Before the poison has any effect Orange kills White. Had Orange not killed White, then the poison would have killed White just a moment later.
> *Overdetermination*. Black and Orange, independently of each other and at the exact same time, shoot White. Each shot is sufficient to kill White.

In Pre-Emption, the full effects of Black's act are interrupted by Orange's act. The effects of Black's poisoning are pre-emptied by Orange's act and White would therefore not have been better off had Black not acted as she did, so Black does not harm White. However, if the details of the case are filled out it could be argued that White would not have been better off had Orange not acted as she did, since if Orange had not killed White then White would have died anyway from the poison. The unintuitive conclusion is that if the Counterfactual Condition is a necessary condition for doing harm then *neither* Black nor Orange harms White. Overdetermination works in a similar way. Here White would not have been better off had Black (or Orange) not acted as she did, so again neither Black nor Orange harms White.

Before considering whether the Weak Counterfactual Condition can be saved from these counterexamples, let us consider what a successful solution to Pre-Emption and Overdetermination would be. In Overdetermination the intuitive verdict is that both Black and Orange harm White. *Someone*, at least, harms White and it would be arbitrary to pick out just one of Black and Orange. In Pre-Emption the intuitive verdict is that Orange harms White. While Black may have done something impermissible we cannot plausibly say that she harmed White because the full consequences of her act do not obtain.

One reply which can be made to Pre-Emption is to claim that White *is* worse off: if Orange had not killed White then White would have been better off until the effects of Black's poisoning come about. From the perspective of White's entire life Orange's act does therefore make White worse off.[35] This reply is hardly satisfying because it is plausible that the magnitude of the harm, the degree by which White is worse off, is the same as the degree of harm Orange does to White. What this reply amounts to is that while Orange harms White, it is not a very serious harm. This problem becomes more acute if we let the effects of Orange's act occur just before the effects of Black's poisoning so that the time between these two events is very small.

It might be objected that this is, on reflection, not as unintuitive as it seems. From a purely prudential perspective it would perhaps not matter so much to

[35] See Feinberg (1986, pp. 152–3).

White if she dies by Black's or Orange's hand if the time between the two events is small.[36] We should not, the objection continues, confuse intuitions about harm and intuitions about morality. Orange's (and Black's) act may be impermissible, or wrong, even though Orange only harms White to a lesser extent than we intuitively think.

This defence does not look very plausible either. Consider the following case from Norcross (2005, p. 166): Black either breaks White's legs or she kills White. If Black chooses to only break White's legs then the Weak Counterfactual Condition would rule out that Black harms White, but this is of course deeply counter-intuitive. Granted, one might think that one should always choose the lesser of two evils, but as before, this seems to be a normative consideration. A defender of the Weak Counterfactual Condition might try to stick to his guns and emphasise that in the morally relevant sense there is no harm in breaking White's legs here. But this suggests that the morally relevant sense of the Weak Counterfactual Condition is so far removed from ordinary uses that it is unclear why one would insist on using harm as a morally relevant notion at all. It should also be noted that the replies just considered only purport to solve Pre-Emption. If the Weak Counterfactual Condition is to be able to reply to Overdetermination in a convincing way then we need a better reply.[37]

#### 2.2.4.1 Collective harm

A trivial observation regarding Overdetermination is that if neither Black nor Orange had shot then White would have been better off. Since both are required to act in a certain way for White's death not to occur, perhaps one should say that Black and Orange together harm White.

It is crucial to this approach to Overdetermination to spell out in what sense Black and Orange can be said to act together. A first suggestion is that they act together in the sense that they perform a collective act. The suggestion would then be that neither Black nor Orange harm White individually but that "they", Black and Orange, harm White.

As Overdetermination is set up the only sense in which "they", Black and Orange, act seems to be a very weak sense: there are two acts, Black's and

[36] Bradley (2012) makes this point.

[37] Note also that appealing to recent developments in the theory of causation will not do to save the Counterfactual Condition. Lewis (2000) suggests an analysis of causation in terms of counterfactual dependence which, he claims, can deal with Overdetermination. On this view, *c* causes *e* if and only if *e* would not have occurred, or would not have occurred in the same way, had *c* not occurred. Assuming this view one could say that had Black not fired his gun then White would not have died in the same way as she actually did. The problem for the Weak Counterfactual Condition is that White would not be better off dying a slightly different way. Death by Black is just as bad for White as death by Black and Orange. As I will argue in chapter five, it seems plausible to analyse the third condition in the basic structure, the responsibility condition, in terms of counterfactual dependence. However, the Counterfactual Condition does not follow from such an analysis.

Orange's, and we can form the "collective" act by simply taking these acts together. The standard view of collective action is however more narrow than this. Intuitively, there is a difference between collective actions and cases where two individual actions contribute to a common outcome. For example, there is a difference between a case where you and I, together, write a paper and a case where you write the first half and I write the second half. A common suggestion is that in the former case there is a kind of collective intention or joint commitment; something which makes the individual acts into a group activity which distinguishes them from a number of mere individual acts.[38]

The current suggestion therefore requires a very wide notion of collective action for it to work as a solution to Overdetermination. This claim, call it "unlimited group composition", has been defended by Tännsjö (1989). Tännsjö suggests that whether we can, or perhaps should, ascribe agency to a group of people depends on "whether we think we have reasons for making moral assessments of the behaviour of groups" (Tännsjö 1989, p.227). Tännsjö notes that if any two acts could in principle constitute a collective action then this would have some unintuitive consequences. However, he adds that even though the collective act consisting of "me writing this chapter and Brutus' killing Caesar" is allowed it does not follow that there is a point to evaluating that collective act. The claim is only that "there are no ontological or methodological reasons against such a classification" (Tännsjö 1989, p. 227).

Unlimited group composition is a strong claim. True, there is a sense of "collective act" in which very disjoint acts constitute a collective but there seems to be a difference between Black and Orange's acts on the one hand and my writing this chapter and Brutus murder of Caesar on the other. What is lacking, it seems, is something which distinguishes Black's and Orange's acts as a collective action in a morally relevant sense while me writing this chapter and Brutus' killing Caesar is not.[39] Without any such difference it would not be warranted to draw any particular conclusions, even *prima facie*, about the moral status of a collective act which satisfies the Counterfactual Condition. The fact that we *can* view Black's and Orange's acts as forming a collective act in the wide sense is not sufficient reason for saying that they acted together in any way that matters. This is not to deny, of course, that we can blame (or praise) acts for their contribution to effects which they are not by themselves sufficient or even necessary for. Nor is it to deny that some groups are more disjoint than others. What we should deny is that Black and Orange constitute a group in the same sense and that they perform a collective

[38] See for example Bratman (1992) and Gilbert (2006).

[39] We could make this challenge more difficult by making Black's and Orange's acts even more disjoint. Suppose that instead of shooting White, Black had in 1654 installed a deadly trap which goes off just when Orange's bullet hits White, and so on. In this way we could make the two collectives more similar and thereby make it more difficult to find any relevant difference.

act. *Pace* Tännsjö, there seem to be reasons against classifying certain sums of individual acts as group acts based on the degree of cooperation, organisation and so forth.

A further difficulty for the collective-act approach to Overdetermination is the assumption that had the collective act, Black's and Orange's, not been performed then neither Black nor Orange would have performed their individual act. But, this is not typically the case for collective acts. For a collective act not to be performed it is sufficient that *one* of the acts constituting the collective act is not performed. For example, *we* (you and I) do not paint a house if I do an you do not. Therefore, the collective act "Black and Orange kills White" is unperformed when Black does not shoot but Orange does (and vice versa).

Finally, it might be suggested that we should view White's death as a synergy effect which emerges from the combination of Black and Orange's act and that White's death cannot be attributed to anything but the collective act. But, White's death is clearly not a synergy effect. Black and Orange's acts are individually *sufficient* for White's death, so how can White's death be something which only emerges from both their acts? As before, we do not have to deny that groups can act, or be responsible, or synergy effects which are not attributable to any particular act. What we should deny, however, is that the collective-act approach is a plausible way to approach Overdetermination.

We should therefore conclude that two acts can overdetermine a state of affairs' obtaining without there being any collective agency. For the collective-act approach to work as a general solution to Overdetermination we therefore have to understand the intuition that "they", Black and Orange in the example above, do harm in a sense which does not rely on Black and Orange performing a collective act.

One suggestion would be that the Weak Counterfactual Condition should be applied to sets of acts and to say that an act does harm only if it is a member of a set which satisfies the condition. However, as the case with me and Brutus shows, not any set will do. A natural suggestion is that the Weak Counterfactual Condition should only be applied to *minimal* sets. For example, "me writing this chapter and Brutus' killing Caesar" is not a minimal set because Brutus' act alone satisfies the condition. In the case of Black and Orange however neither act form a minimal set which White's death is counterfactually dependent on. Together, however, they form a set which is minimal in the intended sense and which satisfies the Weak Counterfactual Condition.

More specifically, we can reformulate the Weak Counterfactual Condition in the following way:

> *The Collective Counterfactual Condition*: an act, $\phi$, harms a person, $b$, only if $\phi$ is a member of a minimal set $M$ such that $b$ is worse off than she would have been had none of the members of $M$ occurred.[40]

[40] See Parfit (1984, p. 71) and Feit (2013).

In ordinary cases, the Collective Counterfactual Condition will have the same implications as the weakened condition. That is, in cases where the minimal set consists only of $\phi$ the new version implies the same thing as the older version. Furthermore, in the same-number cases considered in the previous chapter there is no minimal set such that someone is worse off than they would have been had none of the members of the set occurred. This new version does therefore not solve the non-identity problem and the argument against the Harm Principle still stands. An independent reason for formulating the Counterfactual Condition in collective terms is that we might be able to solve cases where a number of acts together have a harmful effect but where no individual act is sufficient for the effect. For example, suppose ten clever assassins each put a single drop of poison in White's drink. Suppose also that for the poison to have any effect there has to be at least ten drops of poison in White's drink. On the Collective Counterfactual Condition, the ten assassins would do harm because they form a minimal set which is such that White is worse off than she would have been had neither of the set's members occurred.

It is a feature of the Collective Counterfactual Condition, unlike the collective-act view discussed above, that we do not need to claim that Black and Orange perform a collective act. According to the collective Counterfactual Condition, Black's and Orange's acts do harm because they are members of a collective such that had none of the collective's members occurred, a harmful state of affairs (White's death) would not have occurred.

However, formulating the Counterfactual Condition in terms of minimal sets is not enough. Consider the following case:

> *The Death Squads*. Brown is choosing which death squad to join, *A* or *B*. These two groups will then attempt to catch members of The Resistance and execute them. Brown knows that as long as a death squad has at least ten members it will be successful and that group *A* will catch and execute 1000 people while group *B* will catch and execute 10 people. If a squad has fewer than ten members then it will fail to catch any members of The Resistance. Brown also knows that group *A* has ten members and group *B* has nine.[41]

Suppose that Brown wants to minimise the harm he does. It then seems clear that he should not join group *B* because that would push the group to the critical ten-member mark. So he should join group *A*. The Collective Counterfactual Condition implies the opposite however. If Brown joins group *A* then he will be a part of a minimal set consisting of him and any one of the other members of *A* which is such that had these acts not been performed then the 1000 deaths would not occur. That is, he would do a lot of harm. If he joins *B* he would do significantly less harm. According to the Collective Counterfactual Condition, then, Brown should join group *B* if he wants to minimise the amount of harm he does.

[41] This example is an adaptation of Parfit's "Rescue Mission". See Parfit (1984, pp. 67–8).

It could be argued that the difference in the degree of harm which Brown would do makes a critical difference in this example. By adding a condition to the effect that the degree of harm a person does depends on the size of the group he joins one could claim that Brown's contribution is smaller if he joins *A* than if he joins *B* because *A* has more members than *B*. However, note that group *A*, the one with ten members, would catch and kill more members than group *B*. The Collective Counterfactual Condition would therefore still imply that Brown should join *B* rather than *A*.

The question of the degree of harm Brown would do in this case raises a further problem for the Collective Counterfactual Condition. Consider what the view implies regarding Pre-Emption. In this case the smallest group which satisfies the condition consists of Black and Orange. According to this view, Black would harm White in Pre-Emption even though the effects of his act are never realised (because they are pre-emptied by Orange's act). But, this is a rather problematic conclusion because it includes not only acts which actually have a harmful effect but also "backup" acts which would have a harmful effect, had only the circumstances been different.

In defence of the Collective Counterfactual Condition it could be claimed that it is not a good objection that the condition is too wide because it is only a necessary condition for doing harm. Further conditions could be added to rule out that Black harms White in Pre-Emption for example. However, it is unclear what these further conditions would be. According to the Collective Counterfactual Condition, if Orange harms White in Pre-Emption because he is a member of a set which satisfies the condition, then surely Black should as well. Both acts are equally members of the set, so there is no possibility of making a distinction on that ground. Furthermore, if it were to be suggested that Black does not harm White, or does less harm to White than Orange, because Black's act is not related in the right way to White's death, then we seem to have abandoned the collective approach to Overdetermination. If Black's harming White hinges on whether there is the right kind of relation between Black's act and White's death then there simply is no need to appeal to a Collective Counterfactual Condition in the first place.

We should therefore conclude that the Counterfactual Condition is badly situated to answer the objection from Pre-Emption and Overdetermination. The Weak Counterfactual Condition implies that neither Black nor Orange harms White. However, it is clear that at least *someone* harms White. The suggestion that Black and Orange harm White together, but not individually, also fails because it requires an implausible view of collective agency which does away with the distinction between a collective act and a number of individual acts. Applying the Counterfactual Condition to (minimal) collectives also fails. According to this version of the condition there would not be a difference between joining group *A* or *B* in the example of the Death Squads, but there clearly is a difference in this case.

## 2.3 Summary

The conclusion to draw from this discussion is that the Counterfactual Condition is not plausible as a condition for harm in the morally relevant sense. First, in order to defend the Counterfactual Condition against Woodward's objection one would have to commit to a revisionistic view of harm. This is a cost which gives us some reason to see if competing accounts of harm can avoid. Second, the fact that the condition implies that there is no harm done in Overdetermination, or that there might be harm done but not so much in Pre-Emption, gives us strong reasons to reject it as even a necessary condition.

This leaves us with the task of formulating a better alternative. Apart from being faithful to the "pre-analytic core" of harm such an alternative must also save the Harm Principle from the non-identity problem. If there is no better alternative, then the conclusions to draw are rather far-reaching. Suppose that the best analysis includes the Weak Counterfactual Condition. Then it seems that we have to accept some very unintuitive claims about harm, especially with respect to Overdetermination and Pre-Emption. However, it is also an alternative to reject harm as a morally relevant concept. In comparison, this latter alternative would in fact seem more plausible. The consequence, then, of rejecting the Weak Counterfactual Condition is that if there is no better analysis of harm which does not include this condition then we should probably abandon not only the Harm Principle but also the more general view that doing harm is something morally relevant.

# 3. The Non-Comparative View

In the previous chapter I argued that the Counterfactual Condition is not plausible as a necessary nor as a sufficient condition for doing harm. The argument from chapter one against the Harm Principle therefore fails. However, I also claimed that this conclusion is not enough to save the Harm Principle because we still lack an analysis of harm which (i) solves the non-identity problem, (ii) is consistent with the no-difference view and (iii) is intuitively acceptable.

In this chapter I will consider the suggestion that a "non-comparative" analysis of harm achieves these aims. To do harm, according to this view is not a matter of making people worse off but rather to cause people to suffer injuries or other harmful states. This view seems close at hand if we consider what I called "the basic structure" of an analysis of harm. According to the basic structure, to do harm is to make a person be in a harmful state; a state which is non-instrumentally bad for the person who suffers it. Non-comparative analyses of harm emphasise this point. For example, Harris (1990) writes:

> I want to say that to be harmed is to be put in a condition that is harmful. A condition that is harmful [...] is one in which the individual is disabled or suffering in some way or in which their interests or rights are frustrated. [...] I would want to claim that a harmed condition obtains whenever someone is in a disabling or hurtful condition, even though that condition is only marginally disabling and even though it is not possible for that particular individual to avoid the condition in question. (Harris 1990, p. 97).

To do harm, according to Harris, is simply to be responsible for having caused someone to be in a certain "harmful condition". A condition is harmful, furthermore, not in virtue of making the person worse off but in virtue of it having the property of being "disabling" or "hurtful".

Shiffrin (1999) has suggested a similar view. According to Shiffrin, we can "identify harms with certain absolute, noncomparative conditions" (Shiffrin 1999, p. 123) and to do harm is to impose such conditions on a person. Shiffrin goes on to suggest that what unifies the items on this list, what their "harmfulness" consists in, is the following:

> On my view, harm involves conditions that generate a significant chasm or conflict between one's will and one's experience, one's life more broadly understood [...].

> To be harmed primarily involves the imposition of conditions from which the person undergoing them is reasonably alienated or which are strongly at odds with the conditions she would rationally will. (Shiffrin 1999, pp. 123–4).

On Shiffrin's view, to be harmed does not involve any comparison with what would otherwise be the case. To be harmed is rather to suffer a state of a certain kind, a harmful state, and to do harm is simply to cause such states to obtain.

It is not entirely clear whether the views expressed by Harris and Shiffrin are "non-comparative" in any strict sense of the word. For example, Harris' appeals to "disabling" or "hurtful" conditions, but it might be asked whether these notions do not involve a comparison of some kind. Whether a condition is disabling, for example, seems intuitively to involve a comparison of some kind, perhaps with a "normal" condition. Also, as I noted in the previous chapter, harms have a contributive character in the sense that they contribute to a person's overall well-being. We do not merely think that harms are bad for people, we also think that they make life go worse. However, if a "non-comparative" analysis of harm is not supposed to involve any comparison at all then the view expressed by Harris and Shiffrin cannot properly account for this property. If a particular harm contributes to a person's well-being in the sense that that harm makes that person's life go worse (in that respect) then harm-attributions do involve some comparison. Doing harm is therefore not *strictly* non-comparative.

A more plausible way of describing the Non-Comparative View is to say that harm involves *some* comparison, but it is not a comparison with what would otherwise have been the case which is the relevant one. Rather, we should compare with a "baseline situation" which serves as an independent standard for whether an act does harm. The idea, according to this view, is that to harm someone is to make the person worse off in some sense, but not necessarily worse off than the person would otherwise have been. It is still warranted, I think, to call this view "non-comparative" because the way in which harms make life go worse is just the way in which bad things make life go worse. That is, the non-comparative element, that to do harm is to cause a person to suffer a state of affairs which is bad in itself for the person, is primary. The core of this approach can be formulated in the following way:

> *The Non-Comparative View*: an act $\phi$ harms a person $b$ only if $b$ is worse off than she would be in a baseline situation, $S$.

The Non-Comparative View is often claimed to have important consequences for population theory, especially for same-number cases like the non-identity problem. If an act can harm a person even if that person would not have been better off had the act not been performed, then same-number cases could be treated just as same-people cases. Therefore, the non-identity problem would evaporate. The young girl should postpone her pregnancy because if she does

not then she would give her child a bad start in life which, it might be claimed, would be bad for the child. That is, she would *harm* her child if she does not wait. This view also promises to be consistent with the no-difference view. The two medical programmes *A* and *B* would both prevent states which are, intuitively, bad for people so they would both prevent harms.

A common objection which might seem to strike at the Non-Comparative View regardless of how the baseline is characterised is that it does not capture the intuition that preventions of benefits are harms.[1] Suppose that a person is in a condition which is better for her than the baseline and that we can either allow this person to become even better off or to prevent her from becoming better off. It would seem that the Non-Comparative View implies that preventing this person from becoming even better off cannot be to harm her. But, this seems counter-intuitive. As was noted when discussing the Temporal View in the previous chapter, preventing a person's condition from improving is at least sometimes to harm that person.

Another general objection to the Non-Comparative View is that according to this view it is possible to improve a person's condition as much as one can and still harm that person. Suppose White is paralysed from the neck down, and that Black finds a way of improving White's condition so that she is only paralysed from the waist down. White is still in a condition which is bad for her, so Black harms White (and is responsible for the new condition) even though White is significantly better off.[2]

These general objections are not sufficient to refute the Non-Comparative View however. The Non-Comparative View, as I have formulated it, only states a necessary condition for doing harm. It does not follow from the Non-Comparative View that *no* preventions of benefits are harms. For example, preventing a person from becoming better off could be to harm that person according to the Non-Comparative View if the state which the person is in is worse for her than the baseline. Of course, not all preventions of benefits are harms on the Non-Comparative View but as was noted in the previous chapter, it is not a plausible view that all preventions of benefits are harms (recall the kidney-example). Our intuitions here go both ways and it would be premature to rule out the Non-Comparative View simply because it implies that *some* preventions of benefits are not harms.

Both these considerations suggest that the force of this objection will depend on where the baseline is set. We therefore have to take a closer look at specific non-comparative views and see whether they are plausible.

---

[1] See for example Holtug (2002, p. 368), Hanser (2008, p. 430) and Bradley (2012).

[2] See Thomson (2011). Thomson argues that it is a general problem for the Non-Comparative View that some non-comparatively bad states are, intuitively, not harms. However, as I have described the Non-Comparative View Thomson's argument does not work as a general counterexample because the Non-Comparative View only states a necessary condition, not a sufficient condition.

First there is one point about the nature of harm which should be emphasised. What is distinctive about the Non-Comparative View is that to do harm is to make a person be in a bad state. The suggestion is furthermore that the way these bad states contribute to a person's well-being, their contributive character, is captured by a comparison with a baseline rather than what would otherwise have been the case. The point of introducing the baseline is therefore to capture the contributive character of harm. As I will argue below, however, this is a point on which most of the non-comparative views found in the literature fail.

## 3.1 The rational will

As we saw above, Shiffrin suggests that to do harm is to cause a person to be in a condition which "primarily involves the imposition of conditions from which the person undergoing them is reasonably alienated or which are strongly at odds with the conditions she would rationally will" (Shiffrin 1999, p. 124). Disabilities, injuries, illness, pain and death are all harms on this view, she claims, because "[t]hey forcibly impose experiential conditions that are affirmatively contrary to one's will" (ibid.). One way of understanding Shiffrin here is as proposing a form of baseline where the baseline is characterised in terms of what one would rationally will. On this way of reading Shiffrin her view is that to be harmed is to be worse off than one would be in a condition which one would rationally will.

With this view, Shiffrin claims, we can argue that it is possible to harm a person by bringing her into existence, as in the case with the young girl for example. The girl harms her child because she imposes a condition on the child which is contrary to what the child would rationally will. Shiffrin's view also promises to be consistent with the no-difference view. Because her view is formulated in terms of what would be rationally willed it does not seem to involve any comparison with what would otherwise have been the case. If the conditions which we could prevent in the two medical programmes are contrary to what would be rationally willed then it makes no difference that one programme is a same-people case and the other a same-number case.

It is however not clear whether Shiffrin's view delivers what it promises with respect to the non-identity problem or the no-difference view. Consider the young girl for example. If she does not wait she will impose, according to Shiffrin, a condition which is contrary to what that child would rationally will. However, considering the fact that there is no alternative which is better for the child it seems, *pace* Shiffrin, that it would not be irrational of the child to prefer existing with the condition rather than not existing. Since there is no alternative which is better for *this* child, it does not seem irrational for the child to will that she exist with a bad start rather than not exist at all. It is

therefore far from clear that imposing the condition is against what the child would rationally will.

Shiffrin might reply that whether a condition is contrary to what a person would rationally will does not depend on the relational features of the condition. If that is the case then the fact mentioned above, that there is no better alternative for the young girl's child, becomes irrelevant. However, this seems to be an overly harsh restriction on rationality. Surely whether it is rational for me to go to the dentist and undergo a somewhat unpleasant experience there, for example, can depend on what would be the case if I were not to go to the dentist.[3]

A further problem for Shiffrin's view concerns the relation between harm and prudential value. As I argued in the previous chapter, it seems plausible that doing harm requires an effect which is bad for someone. The problem for Shiffrin is that what one would rationally will and well-being can easily come apart; there is simply no necessary connection between these two. What a person would rationally will is not a plausible way to determine whether something is bad for that person because people can have all sorts of odd wills and wants while still remaining rational. For example, it could be rational for me to prefer not to go to the dentist, perhaps because I believe that the experience will be very unpleasant and the expected benefit very small, even though as a matter of fact it would be quite good for me overall. This lack of connection between what one would rationally will and well-being also means that Shiffrin's view fails to capture the contributive character of harm. The fact that a state of affairs is contrary to what one would rationally will does not imply that it is worse for the person to be in such a state. In fact, a state which is contrary to one's rational will can be quite good.

It is possible that Shiffrin has a more substantial view of rationality in mind. The emphasis she places on a person's agency, will and experience suggests more of a Kantian view where "rationality" is not merely understood as coherence among beliefs, desires and other propositional attitudes. To rationally will something should, perhaps, be understood more as "authentic endorsement", or something along those lines.

Understanding Shiffrin's view in this way would however not be much of an improvement. For example, this revised view seems just as vulnerable to counterexamples. Consider a person who endorse oppression or suffering, perhaps because the person believes that she deserves it. We would not say of that person that her endorsement rules out that she is harmed by the oppression or suffering, especially not if she falsely believes that she deserves it. Also, as I argued in the previous chapter, when analysing harm and the way in which harms make life go worse we should not make substantial assumptions about

[3] At one point, Shiffrin claims that "it may be permissible and rational for a person to agree to undergo a harm to receive a benefit, yet it may not be permissible for another party to impose the harm" (Shiffrin 1999, p. 130). It is difficult to see however how such an asymmetrical treatment of doing harm to oneself and doing harm to others can be justified on her view.

well-being. However, this is just what Shiffrin has to do if her view is to have a connection with well-being.

### 3.1.1 Value and value for

So far I have assumed that the Non-Comparative View sets out to analyse harm in terms of what is bad *for* persons. That is, harms are things (in the broadest sense of the term) which contribute in a negative way to a person's overall well-being. An alternative approach, which might be what non-comparativists like Shiffrin have in mind, is to say that we should analyse harm in terms of badness *simpliciter*. The view would in that case be similar to the view once held by Moore:

> In what sense can a thing be good *for me*? It is obvious, if we reflect, that the only thing which can belong to me [...] is something which is good, and not the fact that it is good. When therefore, I talk of anything I get as 'my own good,' I must mean either that the thing I get is good, or that my possessing it is good. [...] In short, when I talk of a thing as 'my own good' all that I can mean is that something which will be exclusively mine, as my own pleasure is mine [...] is also *good absolutely*; or rather that my possession of it is *good absolutely*. (Moore 1993, p. 150).

Reconstructing the Non-Comparative View along these lines might be what Shiffrin has in mind when she identifies harms with certain "evils". Harms would on this view merely be a particular kind of bad events or states, namely those that are possessed or realised by people such as pain, death etc. An episode of pain presupposes that someone suffers the pain and it is therefore a harm rather than just something which is bad in itself. Harms can on this view be contrasted with bad events or states that do not presuppose any person suffering them such as the destruction of beautiful landscapes or inequality.[4]

Identifying prudential value, value-for, with impersonal value seems to be a mistake however. There is no contradiction in saying that a state of affairs could be bad for a person but, intrinsically or finally it is neutral or even good. As has been argued by Sumner (1996), Moore's view seems to get things the wrong way around: "The theory tells us that prudential value depends on ethical [absolute] value: certain conditions make our lives go better because we have a moral reason to bring them about. However, if there is an explanatory relation between ethical and prudential value it seems more likely to run in the opposite direction" (Sumner 1996, p. 51). Sumner's claim is perhaps unnecessarily strong for our purposes. We need not claim that prudential value "explains" ethical value, but it seems correct to insist that there are states which make a life better, or good, while being neutral or even bad from an impersonal point of view. For example, suppose Black and White are applying

[4] Whether there in fact are any bad events which are not also bad for people, in Moore's sense, is not something one needs to take a stand on here.

for the same job. They would benefit to the same degree were they to get it but White has slightly better qualifications than Black. Suppose therefore that it would be best, impersonally, if White got the job. That fact does not invalidate the claim that it would be good *for Black* were he to actually get it. On the view just considered, however, it seems as if we would have to say that it is not actually good for Black to get the job, it might even harm him! But this is obviously quite absurd.

## 3.2 Rights

A different approach to the Non-Comparative View is taken by Woodward (1986). Woodward's view is that the relevant comparison for whether someone has been harmed is with an "(unattainable) baseline situation where [the person] exists and these violations of her rights do not occur" (Woodward 1986, p. 817). This difference between the actual state of a person and the possible baseline state is a difference which "represents a loss which, arguably, one can coherently think of as happening *to* [the person]" (ibid.).

Applied to the non-identity problem, Woodward's view is then that whether the young girl harms her child by not postponing her pregnancy is a matter of whether the child would have been better off had she lived a life where none of her rights were violated. Furthermore, the degree of the harm corresponds to the difference in well-being between these two states.

This view raises more questions than it answers. For example, if rights are a fundamental source of normativity, that is, they generate obligations (perhaps only *pro tanto* or *prima facie*) then it is unclear why harms are morally relevant at all. If it has already been established, as Woodward seems to think, that we violate some right in cases like the young girl, what does it add to say that we also can do harm in such cases? On this view, harm seems to come out as redundant. It should also be noted that as a way of solving the non-identity problem, Woodward's view of harm merely begs the question. The problem with cases like the young girl is to explain why there is a strong reason for her to wait. By saying that the girl would harm her child because she would violate some right possessed by that child is just to assume that there is a strong reason for the girl to wait. Woodward gives no account of how the girl's choice would violate a right possessed by the child.

One way in which harms can be morally relevant in a system of rights is of course by postulating a right not to be harmed. The young girl would violate a right possessed by her child if she does not wait because this choice would harm her child. As should be clear, this is not something Woodward can plausibly claim. If harm is to be analysed in terms of rights then this presupposes that we understand the very notion we are trying to analyse.

Woodward is however a bit unclear as to how central the comparative element is on his view. For example, he claims that

> we think of a person as harmed [...] whenever an action is performed which violates some right possessed by or obligation owed to that person. [...] We thus find it natural to think of the choice of the Risky Policy as harming the nuclear people [...] even though the overall effect of that policy is to leave the nuclear people no worse off than they would be under any possible alternative policy. (Woodward 1986, p. 818).

In this passage, the comparative element drops out and he seems to suggest that harm can be fully analysed in terms of rights violations. However, even without the comparative element Woodward's view seems to be a case of putting the cart before the horse. In the passage above, Woodward assumes a set of rights or obligations which are then used to understand harm. But, it is these very rights and obligations which we are interested in providing a ground for in the first place.[5] A more plausible view is to understand rights partly in terms of harm rather than the other way around.

### 3.2.1 Desert

One way to improve on Woodward's view would be by bringing in the notion of desert. Saying that a person has a right to something sometimes means that that person is entitled to it. Furthermore, the notion of entitlement is closely related to desert; to say that someone deserves something is sometimes to say that that person is entitled to it. If we also suppose that people can deserve to have a certain amount of well-being then we can reformulate Woodward's view in terms of desert: a person has been harmed only if she does not get the amount of well-being she deserves.

This rough characterisation still leaves the view open to many possible interpretations.[6] First, we might mean that a person has been harmed only if she is not as well off as she would be if she got what she deserved. On this interpretation, any discrepancy between a person's actual well-being and her deserved well-being amounts to harm even if the person is better off than she deserves to be. This does not seem like a plausible interpretation because it implies that making someone better off than they deserve would be to harm them. If we take desert to be a morally relevant notion then we might perhaps say that it is not appropriate, in some sense, to make people better off than they deserve but such considerations do not seem to be related to harm in any way.

---

[5] See Holtug (2002, pp. 380–5). Holtug argues that to "moralise" the concept of harm is to make it redundant. By analysing harm in terms of some more fundamental normative concept one is providing an alternative account to the Harm Principle and therefore, in a way, doing away with harm.

[6] See Feldman (1997) who argues for a "desert-adjusted hedonism". For a comment on Feldman and the many ways to understand the idea that people can deserve a certain well-being, see Persson (1997).

On the second way of interpreting the desert-view we might say that a person has been harmed only if she is worse off than she deserves to be. This avoids the problem just mentioned with people being harmed because they are better off than they deserve to be. However, there are several reasons for thinking that this is not a plausible analysis of harm in the morally relevant sense. First, if desert-levels can vary over time then we could harm a person by raising her desert-level. If a person were to become more deserving, by doing good for example, then that person would harm herself by doing good. This seems implausible. Certainly, it would be "appropriate", in a sense, if she were also to become better off in terms off well-being but we would not say that she has harmed herself by doing something which raised her desert-level but not her well-being.[7]

Second, desert and harm seem to be different with respect to their different types of values. As I argued above, it is important to distinguish prudential value, good and bad *for*, from impersonal value, good and bad *simpliciter*. Harms are bad-*for* people, not impersonally bad. Desert, on the other hand, seems to be concerned with value *simpliciter*. It is *good* that a person gets what she deserves, not necessarily good-for that person. Likewise, that a person does not get what she deserves may be, but is not necessarily, bad for that person. If the desert-view was correct then we should not find this discrepancy plausible since on the desert-view if a person does not get what she deserves then this is bad for the person (since it constitutes a harm) and bad *simpliciter* because a person does not get what she deserves.

For example, suppose Yellow is a very talented football player. This talent contributes significantly to making his life well worth living. He enjoys the respect of his peers, the admiration of his fans and so on. Let us also assume that his talent makes him better off than he deserves to be. We might then say that it is inappropriate that he is so well off, and that it might even be better if he was worse off. But it seems very implausible to say that it would not be worse *for him* if he were to become worse off, or that it would not harm him. This suggests that desert and harm come apart and that we should not analyse the latter in terms of the former.

Third, the desert-view only solves cases like the young girl and Depletion if the desert-level is set rather high. If the deserved level of well-being is set low then the view becomes very permissive and would find no objection in the young girl's choice not to postpone her pregnancy. On the other hand, setting the deserved level high makes the view elitistic because creating a person with a life well worth living, but just below the deserved level, would be to harm that person. That is, there would be a reason against creating people even

[7] It might be objected that desert-levels cannot vary over time. People deserve a certain amount of well-being merely in virtue of being people. This reply, it should be noted, departs somewhat from how we usually think of desert. Past actions, for example, are intuitively a ground for assessing what a person deserves. Fortunately, we do not need to resolve this issue since there are further reasons for rejecting the desert-view which are independent of this issue.

though they would have lives well worth living.[8] It is therefore crucial for the desert-view to strike this balance just right, but it is unclear how bright the prospects are for this project. Of course, that it is difficult to specify the desert-level does not show that it cannot be done. However, it raises the question of the explanatory value of the desert-view. The task for the Non-Comparative View was to specify in what sense harms make people worse off. Until we are told more precisely where to place the desert-level the view only amounts to replacing the concept we are trying to analyse, harm, with an at least as poorly understood concept, desert. This new concept, desert, seems to work merely as a placeholder for the right comparison unless the correct desert-level can be specified with at least some precision.

Tying harm to desert in this manner is therefore not a plausible approach. Desert and harm are largely independent of each other. There are of course interesting questions regarding the relation between the two notions, such as whether a person can deserve to be harmed, but these issues no not make it plausible that harm should be analysed in terms of desert.

## 3.3 Health

Harm, as was noted in the introduction to this chapter, is related to concepts such as disability and impairment. An influential analysis of disability and impairment involves the notion of a "normal" state. Whether a condition such as blindness is a disability, for example, is on this view a question of whether it is worse for a person to be in the condition rather than in a normal state.[9] It might be suggested that this analysis can be extended to harmful conditions in general. Harman (2004, 2009) has suggested a version of the Non-Comparative View along these lines. According to Harman:

> An action harms a person if the action causes pain, early death, bodily damage, or deformity to her, even if she would not have existed if the action had not been performed. (Harman 2004, p. 93).

In one respect, Harman's view is very similar to Shiffrin's. Just as Shiffrin she thinks that there are certain states such as bodily damage, deformity and so on which are clear instances of harm. Unlike Shiffrin she makes explicit

[8] The desert-view is similar to so-called "critical level" theories in population axiology. According to such theories there is a critical level of well-being, often placed above where life becomes worth living, such that adding a person to a population only makes it better if the person is better-off compared to the critical level. A common objection to these views is that they imply that it could be better to create a small number of people with lives not worth living than to create a large number of people with lives worth living but below the critical level. For a critical discussion of critical level views in population axiology, see Arrhenius (2000, ch. 5).

[9] This account is not uncontroversial, but it is often used. See for example Daniels (1981), Buchanan (1984) and Buchanan et al. (2001).

the claim that her view involves some comparison. According to Harman the defenders of the Counterfactual Condition mistakenly assume that

> the only available point of comparison is what things would have been like if an action had not been performed. I propose that for persons, there is a point of comparison that involves a healthy bodily state. [...] an action harms someone if it causes the person to be in a state, or endure an event, that is worse than life with a healthy bodily state. (Harman 2004, pp. 96–7).

Harman's view also differs from the Counterfactual Condition, and the Non-Comparative View as I have formulated it, in that she states a *sufficient* rather than *necessary* condition for harming. This is because, as she makes clear, her primary aim is to show that we can harm people in same-number cases and not to develop a complete theory of harm. If causing the kind of state she lists as harmful is a sufficient condition for doing harm then the solution to the non-identity problem is rather straightforward: causing states of the kinds she lists is to do harm and same-number cases are in this respect no different from same-people cases. There is therefore no theoretical obstacle to appealing to harm in these cases.

There are several problems with both her account of harm and her solution to the non-identity problem. First, the assumption that all instances of pain, early death, bodily damage and so on are harms is unwarranted. Typically this is the case, but to conclude from the sensible observation that under normal circumstances pain is a harm that it is always a harm is to ignore the unusual cases. For example, a surgeon might cause bodily damage but we typically do not think that the surgeon harms her patient. Being in one of these conditions is therefore not sufficient for being harmed. A consequence of this is that her solution to the non-identity problem fails. That solution relied on the claim that being in a state of a certain kind is sufficient to be harmed, but that claim seems questionable and Harman gives us no reason to believe it. Second, Harman's focus on a healthy *bodily* state makes her view very narrow. If a person is in a healthy bodily state, but is suffering mentally then that person would not, it seems, be harmed according to Harman's view.

In reply to this objection it could be suggested that we should modify Harman's view. Health, it might be agreed, is too narrow but we can remedy this deficit by formulating her view in terms of "normal functioning" instead. Paradigmatic examples of harmful states such as disability, disease and deformity are often analysed in terms of normal functioning. This suggestion would then amount to the claim that what these paradigmatic examples have in common is that they are worse for their victims than a life with normal functioning and that this is true of harms in general.[10]

[10] Buchanan et al. (2001, ch. 3) seem to endorse a similar view. However, it is unclear whether they claim that it is only *disabilities* which should be defined via what is "normal" or whether this is true of harm as well. For a critical discussion of the significance of normality to disabili-

However, we should reject this revised version of Harman's view as well. This view implies that some treatments which we would typically count as beneficial to a person are in fact harmful. Suppose, for example, that we can only partially cure a person's blindness. Being partially blind, let us assume, is better for the person than total blindness, but worse for the person than having normal vision.[11] According to the normal functioning view, and Harman's view, we would *harm* this person by partially curing her blindness. But this seems quite implausible.

Furthermore, analysing harm in terms of normal functioning, or health, does not capture the way in which a finally bad state of affairs makes life go worse because there does not seem to be any close connection between a person's well-being and normal functioning. For example, suppose that being near-sighted is worse for a person than having normal vision. Even though we assume that this person's life would go better if she had normal vision, it does not seem to follow from that fact that she is in a harmful condition, and that her lack of normal vision makes her life go worse in the relevant sense. Of course, it makes her life go worse than it would go if she had normal vision, but this simply does not seem to be relevant when it comes to the question whether her near-sightedness is a harmful state. In order for the normal functioning view to be a plausible analysis of harm one would have to make substantial assumptions about well-being, namely that there is a connection between well-being and normal functioning. But, an analysis of harm should not restrict the available theories of well-being in this way.

## 3.4 Summary

None of the non-comparative views which I have considered in this chapter are plausible as analyses of harm. There are two general lessons to draw from these views however. First, as we saw with the Moorean view and the desert view, harm should be analysed in terms of prudential value and not impersonal value. To harm someone is to make that person's life go worse in some sense. Whether harms *also* have impersonal value is a further claim which is independent of the relation harms bear to individual well-being. This should not be much of a surprise, and not very controversial, but it is worth emphasising. Second, the contributive value of harm is not plausibly captured by using a "baseline" as a comparison. At least, none of the baselines we have considered in this chapter are plausible without making substantial assumptions about well-being.

---

ties, and to harm, see Kahane & Savulescu (2008, 2012). They argue that normality, perceived as either biological normality or statistical normality, is not intrinsically relevant to disability or harm though the latter can be relevant derivatively.

[11] Thomson (2011, p. 441) raises this objection as a general argument against the Non-Comparative View.

# 4. The contributive value of harm

When analysing harm it is important to distinguish between harm in the final sense and in the instrumental sense. To rehearse this distinction, final harms are states of affairs which are bad for people because of their nature while instrumental harms are states or events which are harmful because of their effects. In chapter two it was suggested that acts, in so far as they are harmful, are typically instrumentally harmful and that a plausible analysis of instrumental harm should be in terms of final harm. I also introduced three conditions which belong to a pre-analytic core and which an analysis of harm should preserve:

> *a* harms *b* only if
> (1) *a* performs an act, $\phi$,
> (2) *b* is in a state *S* which is bad for *b* in the final sense.
> (3) $\phi$ is responsible for *S*'s obtaining.

According to this basic structure, a necessary condition for doing harm is that there is a state of a certain kind, viz., a state of affairs which is finally bad for a person. It was suggested in chapter two that we can think of these states of affairs as components of a person's well-being. In the previous chapter I argued that The Non-Comparative View preserves this claim but it fails to capture the plausible claim that harming a person is to make that person's life go worse. This is also true of the basic structure. None of the conditions (1)-(3) captures the the contributive character of harm. Accounting for this feature of harm is the main task of this chapter.

A plausible starting point is to consider whether this can be done by revising the second condition in the basic structure. If there is a connection between negative well-being components, states which are bad for people, and the contributive value of such components then one could capture the way in which harms make life go worse by spelling out the contributive value of finally bad states of affairs.

In this and the following two chapters I will make an attempt at analysing harm by building on this basic structure. I will have very little to say about (1). In the next chapter I will discuss how (3) can be further analysed in terms of counterfactual dependence and in chapter six I will argue that we should not add any further conditions to the basic structure pertaining to intentions, foresight or consent. The focus of this chapter will be on (2) and in what way

a state of affairs which is bad for a person in the final sense, a negative well-being component, makes a difference to a person's well-being.

According to the view I will argue for, *the Same-World View*, a state of affairs has negative contributive value for a person if and only if that person's life-time well-being is lower when we take the state of affairs into account than it is when we do not take the state of affairs into account. The details of this view will be spelled out below. I will also argue that we should modify the second condition in the basic structure. Depending on what the correct theory of well-being is, it is possible that a state of affairs which is neutral, or even good, for the person has negative contributive value. I will argue that what matters to doing harm is not that there is a state of affairs which is bad in the final sense for a person, but rather that there is a state of affairs which has negative contributory value for a person. I will thus suggest that we should replace (2) with:

> (2*) *b* is in a state *S* such that *b*'s life is worse, taking *S* into account, than *b*'s life not taking *S* into account.

By replacing (2) with (2*) in the basic structure, it is not necessarily the case that acts are instrumentally harmful because they have effects which are finally bad for a person. I will argue that there are two routes by which one can arrive at (2*) as a replacement for (2). First, one can claim that it is actually the contributive value of a state of affairs which is relevant to harm, not the final value. Second, if all finally bad states of affairs have *some* negative contributive value, and all finally good states of affairs have *some* positive contributory value, then (2) implies (2*).

Finally, I will consider what the Same-World View about contributive value, and (2*), suggests with respect to puzzling cases such as the value of existence and the non-identity problem.

## 4.1 The Simple View

Let's start with a simple idea. A classic analysis of the relation between goodness and betterness is that a state of affairs is *good* if and only if it is better than its negation and a state of affairs is *bad* if and only if it is worse than its negation.[1] Perhaps this traditional view, suitably modified, can serve as a basis for understanding the relation between negative well-being components and their contributive value.

The traditional view is not a substantial theory of goodness. That is, it does not tell us what actually *is* good and bad. The traditional view is rather a

[1] An early formulation of the simple view is Brogan (1919). The claim that states of affairs are the bearers of value is commonly assumed, see for example Chisholm (1968, pp. 22–3). For a more pluralistic view about the bearers of value, see Rabinowicz & Rønnow-Rasmussen (2000, pp. 46–7).

formal requirement which should be compatible with any substantial view about goodness. This neutrality with respect to substantial theories can also be applied to well-being and to the task of spelling out the contributory value of a state of affairs *for a person*. What we are after is a way of expressing the relation between bad-for and worse-for without making any assumptions about what is bad for a person.[2]

The traditional view has, however, been subject to some serious criticism. Chisholm & Sosa (1966) has argued that the traditional view is incompatible with some very plausible axiologies. For example, according to hedonism pleasure is the only intrinsic good and pain the only intrinsic bad. Now consider the state of affairs "there are no unhappy egrets". This state of affairs seems to be better than its negation ("there are unhappy egrets") assuming hedonism. However, it is not good according to hedonism because it does not involve any pleasure. It is, as they say, not a state of affairs which would "rate any possible universe a plus" (Chisholm & Sosa 1966, p. 245).[3] That the traditional view rules out this form of hedonism is a serious blow if the traditional view is to be about the logic of value and has lead many philosophers to reject the this view.[4]

As the objection is formulated, it targets an attempt to reduce *good* to *better* rather than *good for* to *better for*. However, the objection could easily be extended to the latter as well. Consider for example hedonism as a theory about prudential value rather than impersonal value. That is, pleasure is the sole good-for and pain the only bad-for. Now consider the state of affairs "*b* is not in pain (at time *t*)". According to hedonism, this state of affairs is not good for *b* because it does not involve any pleasure, only the absence of pain. That is, it does not rate *b*'s *life* a plus (or a minus) from a purely prudential perspective according to hedonism. However, that "*b* is not in pain (at time *t*)" is clearly better than its negation for *b*, so the traditional view implies that it is good for *b*, assuming hedonism.

However, note that this objection to the traditional view only concerns the right to left implication. That is, the state of affairs "*b* is not in pain (at time *t*)" is better than its negation and is therefore, according to the traditional view,

---

[2] In the literature it is common to make a tripartite distinction between theories of well-being into hedonism, desire-satisfactionism and objective-list theories. See for example Parfit (1984, appendix I) and Sumner (1996, chs. 3–5). My examples will often be formulated in terms of hedonism, but this is merely to simplify the exposition.

[3] Chisholm (1968, p. 24) suggest a slightly different reason for saying that the state of affairs "there are no unhappy egrets" is not good. Here Chisholm claims that the reason why "there are no unhappy egrets" is not a bad state of affairs is because it does not entail any state of affairs which is intrinsically bad. In contrast, the state of affairs "there *are* unhappy egrets" is bad because it entails an, arguably, intrinsically bad state of affairs. This argument seems to be question begging however. Since every proposition entails itself, it has to be assumed that "there are no unhappy egrets" is not intrinsically good. But this claim is just what is at stake here. Whether Chisholm's criterion, if it should be understood as a criterion, is plausible is a question I will not address here.

[4] See Chisholm & Sosa (1966) and Åkvist (1968).

good for *b*. But since we would not say that it is good we have to deny the conditional "if a state of affairs is better than its negation then it is good". The left to right implication is however not affected by this counterexample. We can agree that the absence of an episode of pain is not in itself good for a person but still claim that everything that is good for a person is better than its negation for the person. Furthermore, the present task is to formulate in what sense a bad state of affairs, a final harm, makes a person's life go worse. Fulfilling this task only requires the weaker claim that if a state of affairs is bad for a person then it is worse than its negation. By formulating a weaker version of the traditional view in terms of prudential value, value-for, we get a first attempt at capturing the contributive value of harm:

*The Simple View*: if *S* is finally bad for *b* then *S* is worse for *b* than ¬*S*.

The Simple View satisfies the neutrality requirement mentioned above. It does not commit us to any particular theory of well-being. It also has some heuristic uses. When testing a theory of well-being against our intuitions it is possible to argue, via the Simple View, for a theory of well-being by showing that what the theory says about what is bad for a person coheres with intuitions about what is worse for a person. Likewise, one can argue against a theory of well-being by showing that what the theory claims with respect to what is worse for a person is incompatible with our intuitions about what is bad for a person. For example, hedonism as an example again, the view that pleasure is good for people but that the presence of pleasure is not better for people than the absence of pleasure would be ruled out by this view.

As a conceptual claim about final prudential value the Simple View seems plausible. It seems very plausible that, as the Simple View claims, all negative well-being components are worse for people than their negation. However, as a claim about the sense in which a bad state of affairs makes life go worse it does not seem to be sufficient. In order to capture the contribution a state of affairs makes to a person's well-being it seems necessary to take the circumstances of that particular person's life into account. For example, "being half-blind" would (probably) make a negative contribution to the life of a person with normal vision but could make a positive contribution to a blind person's life. The Simple View is therefore not a view about the contributive character of harms. Rather, it is a view about how different well-being components relate to each other and not a view about how a negative well-being component is related to contributive value. But this is something else than the contribution of a state of affairs to how well a person's life goes. What is missing is a claim about the contribution a bad state of affairs makes to the value of a particular life for a person.

### 4.1.1 Subtraction and replacement

An alternative to the Simple View has been suggested by Arrhenius (2013, pp. 29–35). On his view we should define intrinsic value in terms of *neutral* value.[5] With the notion of a neutral state of affairs it is then possible to formulate the relation to comparative prudential value in the following way:

> *The Neutrality View*: If a state of affairs *S* is finally bad for *b* then *S* is worse for *b* than a state of affairs which is neutral for *b*.

The Neutrality View does not seem to be a view about contributory value either. As with the Simple View, the Neutrality View seems plausible as a conceptual claim about the final value of well-being components but it does not address the contributory value of negative well-being components. However, Arrhenius (2013, p. 34) suggests that we can analyse contributory value in terms of neutrality in the following way:

> *Neutral Contribution*: *S* is neutral relative to a certain life *x* if and only if *x* with *S* has the same well-being as *x* without *S*.[6]

This claim seems to be on the right track. It relates the contributory value of a state of affairs to a particular life while remaining neutral with respect to substantial theories of well-being. That is, Neutral Contribution is compatible with *S* being neutral relative to your life but not neutral relative to mine. Furthermore, Neutral Contribution is also compatible with the view that *S* can be neutral in itself but good (bad) relative to a particular life. As an example, consider states of affairs which are not valuable in themselves but which are preconditions for other states of affairs which are valuable in themselves. Consciousness, according to hedonism at least, is a precondition for good and bad states of affairs but is neutral in itself. However, consciousness can have non-neutral contributive value according to Neutral Contribution.

Neutral Contribution is then plausibly seen as an analysis of neutral contributive value. It does not, without further assumptions, tell us how negative contribution should be understood and in what way a bad state of affairs makes a life go worse. One way to extend the view to cover negative contribution would be to analyse negative contribution in terms of neutral contribution:

> *Negative Contribution-1*: *S* is bad relative to a life *x* if and only if *x* with *S* has lower well-being than a life with a state of affairs which is neutral relative to *x*.

Alternatively, we can analyse negative contribution in a similar way as we analysed neutral contribution:

---

[5] Arrhenius' view can be seen as an adaption of the analysis of intrinsic, impersonal value found in Chisholm & Sosa (1966).

[6] This claim corresponds to Arrhenius' (***). He uses the term "welfare" instead of "well-being" when formulating (***). This is merely a difference in terminology.

> *Negative Contribution-2*: *S* is bad relative to a certain life *x* if and only if *x* with *S* has lower well-being than *x* without *S*.

The main difference between the two seems to be that in Negative Contribution-1 we *replace* the state of affairs *S* with a state of affairs which is neutral and then compare these two possible lives with respect to how much well-being they contain. In Negative Contribution-2 we *subtract* the state of affairs from a life and then compare the two lives with respect to their well-being.

A problem for subtraction is however that it is unclear exactly how we are to understand the comparison of a life "without" a state of affairs. As I just noted, one way to understand a life without a state of affairs is as merely subtracting the state of affairs from that life. This seems conceivable when it comes to some states, for example particular experiences of pleasure or pain. It is much more unclear if we consider for example consciousness, or the state of affairs that a particular person exists. If we are to only subtract the state of affairs that a person is conscious, then the subtraction-strategy instructs us to compare a life with consciousness and all the mental states it contains with a life without consciousness but with the same mental states. But this is hardly conceivable.

This problem for the subtraction-strategy suggests that we should replace rather then subtract. However, a similar objection can be raised against the replacement strategy. For some states of affairs it is unclear what we should replace these states of affairs with. For example, if we are to replace the state of affairs "*b* exists", what should we replace it with? This problem is not very surprising because the replacement strategy involves subtraction. On the replacement approach we are to consider a life with a state of affairs and compare it with a life without *that* state of affairs but with a neutral state of affairs instead. In so far as there is a problem involved with subtraction, it is a problem which seems to strike against the replacement strategy as well.

However, there is one difference between the two strategies which suggests that the replacement strategy is better positioned to solve this problem than the subtraction strategy. The difference is that the subtraction strategy seems to involve inconceivable scenarios, for example that a person has mental states but is not conscious (because we "subtract" the state of affairs that she is conscious).[7] The replacement strategy, on the other hand, could perhaps be spelled out in such a way so that we avoid this conclusion. In the next section I will consider just such a take on the replacement strategy: the Similarity View.

[7] Not all subtractions need result in inconceivable scenarios of course but when it comes to states of affairs like "*b* exists" or "*b* is conscious" this might be the case.

## 4.2 The Similarity View

A suggestion which could be made at this juncture is that we can spell out the "replacement-strategy" mentioned above by using the terminology of "possible worlds". Possible worlds are common in philosophy and I intend my use here to be non-committal with respect to how they should be understood more precisely. As I will use the term, a possible world is a way in which the world might be and which propositions may be true or false relative to. Saying that a state of affairs obtains in a possible world is just to say that a certain proposition is true relative to that world.

Using possible worlds we can understand the replacement strategy as saying that when we replace a state of affairs we should look at the most similar possible world where that state of affairs does not obtain. A finally bad state of affairs, on this view, makes a particular life go worse because the world where the state of affairs obtains is worse for the person than the most similar possible world where it does not obtain. That is:

> *The Similarity View*: if *S* is bad for *b* then the actual world where *S* obtains is worse for *b* than the most similar possible world where *S* does not obtain.

A problem with the Similarity View concerns how to determine the closest possible world where *S* does not obtain. McMahan (1988, pp. 46–7) has argued that in order to specify what would be the case were *S* not to obtain we should employ so-called backtracking.[8] McMahan's idea, roughly put, is that the most similar world where *S* is not the case is a world where neither *S* nor "the entire causal sequence of which the immediate cause of [*S*] is a part" (McMahan 1988, p. 47). For example, if the state of affairs *S* is someone having a broken leg then we should look at the history of the actual world, identify what caused this state of affairs, and then make the smallest possible change which would result in that person not having a broken leg.

Using backtracking, the Similarity View would be very similar to, if not equivalent with, the Counterfactual Condition. As a consequence, we would not harm people by, for example, creating them with physical or mental impairments. Consider the example of the young girl again. The putative harm in this case is the child's "bad start in life". On the backtracking interpretation of the Similarity View it seems that we should in this case compare with the possible world where the girl decides to postpone her pregnancy because this is, arguably, the smallest historical change which is sufficient for the state of affairs "the child has a bad start in life" not to obtain. However, this is a possible world where the child does not exist and therefore, for the reasons given in chapter two, it is not worse for the child to have a bad start in life.[9]

[8] The term "backtracking" is from Lewis (1979).

[9] See for example Bayles (1976, p. 296) and Thomson (2011, pp. 446–7). Thomson argues that the Similarity View fails because of "reasons familiar to us from the failure of the simple counterfactual analysis of causation" (p. 446). Here she seems to assume that the most similar

However, it would be a mistake to use backtracking in this context. As a claim about the contributive value of *final* harms, states of affairs which are bad for people because of their nature, the backtracking interpretation of the Similarity View allows for too large differences between the *S* world and the ¬*S* world. If the cause of *S* is necessary for other effects then to compare with a world where the cause does not occur will of course be to compare with a world where these other effects do not occur either. This is not what we want when we ask whether a particular state of affairs is bad as an end for a person. What we want from such an analysis is whether the state of affairs in question makes a difference to someone's well-being. With the backtracking interpretation it will not always be possible to single out the difference *S* makes on its own since the world we are comparing with differs in so many other respects. This is not to deny that backtracking might be the proper way to evaluate ordinary counterfactuals. The claim is simply that we should not, in the present context, employ an ordinary counterfactual but rather a "tailored" one where it is specified which states of affairs we should keep fixed.[10]

In order to determine whether the Similarity View is a plausible way of spelling out the contributive value of final harms we should therefore consider how we might "tailor" the counterfactual more closely. This would involve identifying those factors which we allow to vary and those which should be kept fixed when determining what the most similar world is. One such factor which seems clearly relevant to the similarity of two worlds is to what extent they share the same states of affairs. For example, if Obama wins the presidential election of 2008 in world $w_1$ and $w_2$, but not in $w_3$, then $w_1$ and $w_2$ are more similar to each other than they are to $w_3$, other things being equal. A second factor which seems highly relevant is whether two worlds obey the same laws. If two worlds differ only with respect to one state of affairs then one of them will be very peculiar since the history in that world, what goes on after the state of affairs obtained, is the same as it is in the world where the state of affairs does not obtain. For example, if $w_1$ and $w_3$ only differ with respect to the state of affairs "Obama wins the 2008 election" then it would still be the case in both worlds, even the one in which he loses, that Obama is president in 2009, that he introduces Obamacare and so on.

So far we have noted sameness of states of affairs prior to *S*'s obtaining and laws as relevant to similarity. Saying that these two are relevant leaves it open how they interact. Sameness of states of affairs prior to *S*'s obtaining

world is determined by backtracking, but as we will see that is not the only option available to the Similarity View.

[10] A further reason to think that the backtracking view is not plausible in this context is that what the most similar world is in a case like the young girl might depend on the context of the counterfactual. For example, if we are interested in what would have been the case had the girl's child not had a *bad* start, then it seems plausible to say that the most similar possible world is one where the same child exists but with a different start in life. However, if we are interested in what would have been the case had *her child* not had a bad start, then it seems plausible that the most similar possible world need not be one where the same child exists.

seems to be more important to the similarity of two worlds than sameness of states of affairs after *S*'s obtaining. What the example with Obama illustrates is that the most similar world where Obama does not win the 2008 election is a world where the states of affairs prior to the election are the same but where the states of affairs which obtain after the election are allowed to differ. How these two factors, sameness of states of affairs prior to *S*'s obtaining and sameness of laws, interact should they come into conflict is still unclear and leaves a lot of room for interpretation.

One response to the consequence that the similarity relation leaves a lot of room for interpretation is to endorse it. Bradley (2004) for example, says:

> What counts as the most similar world to the actual world is not always a determinate matter. It depends on what features of the actual world we want to keep fixed, or on what similarity relation we are employing. [...] We cannot get a determinate answer to the question 'what would have happened?' until we decide what must remain fixed and what may vary. (Bradley 2004, pp. 49–50).

Bradley seems here to be following Lewis (1986, p. 21) in claiming that which departures we should accept depends on the context of the counterfactual and what the purpose of evaluating it is. To take one of Lewis' examples, if we are asking what would be the case if kangaroos had no tails we generally ignore worlds where kangaroos "float around like balloons". In asking what would be the case if kangaroos had no tails then we are not interested in worlds where kangaroos have such, relative to our world, additional floating properties.

Regarding the contributive value of finally bad states of affairs, the question what we should keep fixed and what should be allowed can partly, at least, be answered by considering what would be gratuitous departures with respect to the question 'is *S* bad for *b*?'. One thing which would constitute a gratuitous departure would be any change to the components of *b*'s well-being which are distinct from *S*. That is, all and only those states of affairs which are components of *b*'s well-being in the *S*-world and which are distinct from *S* should also obtain in the $\neg S$ world. If we were to allow these to vary then the comparison between how well off *b* is in the *S*-world compared with the $\neg S$ world would not tell us the value of *S* for *b*, but the value of *S* and something else.

We now seem to have arrived at two plausible claims about the similarity of worlds which the Similarity View should satisfy. But, there is an obvious tension between these two claims. First, there is the claim just made that the most similar world to the actual one where *S* does not obtain should be identical with respect to well-being components distinct from *S*. Otherwise, the Similarity View would not be a view about the contributive value of *S* only. This means that any well-being component obtaining after *S* in the actual world should also obtain in the most similar world. Second, we should allow the most similar $\neg S$ world to differ at times after *S*'s (non-) obtaining. As

was noted when discussing backtracking however, worlds which "converge" in this manner are very dissimilar because they would not obey the same laws.

Resolving this apparent conflict by giving up either claim is of course an option but not a very plausible strategy. Giving up the claim that we should allow the $\neg S$-world to differ from the actual at times after $S$ is not plausible because the laws of such a world would be very different from the actual world. Giving up the claim that we should keep distinct components of $b$'s well-being fixed also seems implausible because the Similarity View would then not be a claim about the value $S$ has for $b$ in itself but rather the value of $S$ and something else for $b$.

### 4.2.1 Time as a factor

As we saw above, a problem for the Similarity View is that there is an apparent conflict between on the one hand capturing the contributive value of $S$ for $b$, and not the value of $S$-and-something-else for $b$, and on the other hand the claim that the most similar $\neg S$ world to any given $S$ world should be allowed to differ from the $S$ world. In order to resolve this conflict we should consider an assumption which has so far been tacit in the discussion, namely that when comparing the $S$ world with the most similar $\neg S$ world we are comparing how well $b$'s *life* goes in the two worlds. An alternative to comparing whole lives is to compare a part of $b$'s actual life with a part of $b$'s life in the most similar $\neg S$ world. The two restrictions seem to conflict when it comes to states of affairs which obtain after $S$, but if we simply ignore what goes on after $S$ in the actual as well as the counterfactual world then that should not be an issue.

One straightforward way to amend the Similarity View would be to say that we compare the part of $b$'s actual life up to and including the time $t$ when $S$ obtains with how $b$'s life would have gone up to, but not including $t$, in the most similar world where $S$ does not obtain. This does solve the problem of taking into account what goes on after $S$, although it is still not a plausible condition of when a *state of affairs* is bad for a person. Rather, this seems to give a sufficient condition for when a *time* is bad for a person. Many distinct states of affairs obtain at $t$, and a life which includes $t$ can differ from a life which does not in many respects other than the obtaining of $S$.

One way to avoid this problem is to consider parts of a person's life of equal length:

> *The Time-Relative Similarity View*: if a state of affairs $S$ obtaining at time $t$ is finally bad for $b$ then $b$'s life up to, and including, $t$ is worse for $b$ than $b$'s life up to, and including, $t$ in the most similar possible world where $S$ does not obtain.[11]

[11] The time-variable '$t$' could of course stand for an interval rather than a point in time.

On the Time-Relative Similarity View, the most similar $\neg S$ world can differ in all sorts of respects at times after $t$ without affecting the contributive value of $S$ for $b$. Also note that although this view refers to time it does not imply anything about *when* $S$ is bad for $b$. It has been argued that a person can be harmed by states of affairs which obtain after her death and that such posthumous harms are bad for the person while she is alive.[12] The Time-Relative Similarity View is only a sufficient condition and does therefore not rule out that posthumous harms can affect a person's well-being at times before they obtain. What the view commits one to is the claim that a world including a posthumous harm, that is, a world up to and including the time when the harm obtains, is worse for the person than the same time span in the most similar world where it does not obtain.[13]

A problem for this view, and which strikes against the previous versions of the Similarity View as well, are cases where a state of affairs obtains which is *prima facie* bad for someone, but where there is a back-up which ensures that the person would not have been better off had this state of affairs not obtained. Consider the following example:

> *The Back-Up*: Black and Orange have been hired to kill White. Black has been instructed to shoot White while Orange has been instructed to detonate a bomb which will kill White if and only if Black does not shoot. Black shoots and kills White.[14]

In this case, had White not died from Black's shot she would still have died from the explosion triggered by Orange. It is therefore not the case that she is worse off in the most similar possible world where her death, $S$, does not obtain because it is not worse for White to be killed by Black than it is to be killed by Orange. Therefore, the Time-Relative Similarity View implies that White's death cannot be bad for her.

---

[12] See Pitcher (1984), Feinberg (1987, pp. 83–93) and Luper (2004) for example.

[13] A more problematic case concerns states which are *only* good or bad for a person before they obtain. For example, it could be claimed that the satisfaction of past preferences are good for people even if the preference is not held when it is satisfied. According to this theory, a state of affairs can be good for a person before it obtains but neutral for the person when it obtains (because the person does not then have the relevant preferences). It is unclear, however, whether this is a coherent preferentialist theory. If it is the *satisfaction* of preferences that matters then it seems, *prima facie* at least, that a preference which is not held when its object obtains does not count as satisfied, and does not make a person's life any better. If it is the *object* of a preference which carries the value, then there seems to be no reason to say that the object of a preference is good for a person before the object obtains. See Rabinowicz & Österberg (1996) regarding the distinction between "object" and "satisfaction" interpretations of preferentialism. See also Bykvist (1998).

[14] This case is obviously similar to the overdetermination-example discussed in chapter two. The difference is that in chapter two we were concerned with the the question whether Black and/or Orange harm White. Here we are concerned with whether White's death is bad for her according to the Time-Relative Similarity View.

Bradley (2004, p. 54) has argued that this problem can be solved if we consider the vagueness of what we are allowed to vary when determining what the most similar possible world is like. Bradley suggests that, depending on the context of evaluation, the most similar possible world can be one where White does not die at all. For example, if we are asking "was White's death due to Black firing his gun bad for her, given that Orange would have detonated the bomb had Black not fired his gun?" then the answer, according to Bradley, must be no. However, if we are asking "was it bad for White to die when she in fact did, rather than not?" then the question has been specified so that we can rule out the worlds where Black shoots and the worlds where Orange detonates the bomb as the most similar one.

A similar view is suggested by Feldman (1991) who argues that there are many states of affairs involving White's death and each of these should be evaluated independently. For example, the states of affairs "White is killed by Black" and "White dies at exactly time *t*" are perhaps not bad for White but the state of affairs and "White dies" plausibly is. If "White dies" does not obtain, then the most similar possible world is one where neither Black nor Orange kill her. Since this state of affairs is worse for White, the Time-Relative Similarity View does not rule out that it is a bad thing for White to die in this situation.

Bradley and Feldman attempt to specify the most similar world in such a way that neither Black nor Orange kill White in the most similar world. However, it is not clear that the specifications are sufficient for it to follow that White's death is bad for her if we consider other factors which are necessary for White to be better off in the most similar world where neither Black nor Orange act. The problem, it seems, is that we still have to replace the state of affairs which is bad for White with something else, and unless the replacement-state is specified enough to rule out that White does not die, or suffers a fate worse than death, we cannot conclude that White's death is bad for her.

The obvious reply to this objection is perhaps that the only thing the objection shows is that we need to specify the circumstances of evaluating White's death more. On Bradley's view, for example, we can perhaps specify the relevant similarity relation further so that the closest world is one where neither Black nor Orange act, and where she does not suffer a fate worse than death. Similarly, on Feldman's view one could perhaps also argue that there is a more complicated state of affairs involving White's death which is bad for her.

I will not pursue the Similarity View further. It is surprising, I think, that a case like Back-Up is even a problem for a view about contributive value. Intuitively, it seems to me, the fact that White would have died in a different way had she not died by Black's shot is irrelevant to whether White's death is bad for her and whether White's death makes her life go worse. It is, of course, relevant to whether it would have been better for White not to be killed by Black but that is a different matter. It would not have been better for White

not to be killed by Black, but that is because something bad happens to White, she dies, and something equally bad would have happened had she not died.

Another reason not to pursue the Similarity View is that there is a more straightforward and simpler approach which does not require a similarity relation nor the metaphysics of possible world. On this view, which I will suggest below, one does not compare how well off a person is in different possible worlds but simply looks at one world. To introduce possible worlds seems to have brought more problems with it than it has solved and it would be a significant advantage if we can do without the similarity relation and the metaphysics of possible worlds.

## 4.3 Same-world comparisons

A persistent problem for the Similarity View is how to determine what the most similar world looks like. As I argued above, the notion of "similarity" involves a considerable amount of vagueness and it would be an advantage if we could remove this vagueness. The Similarity View also faced the problem of taking irrelevant considerations into account. If the most similar world where $S$ does not obtain differs in other respects than with respect to $S$, then the difference in value for a person between the $S$ and the most similar $\neg S$ world would be the difference between $S$ and these other respects, not just the value of $S$. We also saw that putting restrictions on the non-$S$ world, for example with respect to time, did not completely remove this problem.

In order to formulate a view which avoids these problems we can start with the observation that the most similar world to any world is, trivially, that world itself. If we can spell out the contributive character of a particular state of affairs by considering only the world where the state of affairs obtains, then we would not have the problem of vagueness and irrelevant considerations. However, this might seem like an impossible task since, if we are to build on the Simple View, then the contributive character of a state of affairs involves a comparison between the state of affairs and the "negation" of a state of affairs. So it might seem that we have to compare with some other world, the non-$S$ world.

What I suggest is that we should reinterpret the Simple View in the following way:

> *The Same-World View*: if a state of affairs $S$ is finally bad for $b$ then $b$'s life, taking $S$ into account, is worse than $b$'s life not taking $S$ into account.

On this view, that $S$ is worse than not-$S$ for $b$ does not mean that the possible world where $S$ does not obtain is worse for $b$, but that $b$ has less well-being if we take $S$ into account than if we do not. This view involves a same-world comparison in the sense that we only look at the world where $S$ obtains and

then ask how well-off $b$ is if we count $S$ and how well off $b$ is if we do not count $S$. On the same-world view we do not replace $S$ with some other state of affairs, that which would be the case in the most similar possible world for example. Instead, we merely "subtract" $S$ by not taking it into account.

What exactly does it mean to "take $S$ into account"? Here we need to introduce the notion of a well-being *function*. A well-being function is an operation which takes the well-being components that obtain in a life as input and generates the overall value of that life for the person as its output. To take a state of affairs into account is then for the well-being function to take the state of affairs as a value.

Put in a more formal way, let $V$ be the set of all the well-being components that obtain in $b$'s life. The value of $b$'s life for $b$ is a function of $V$, $f(V)$, and the same world view is then merely the claim that if a member of $V$, $S_i$, is bad for $b$ then $f(V) > f(V - S_i)$.[15] An example of a well-being function is to simply add the value of each component so that the value of a life is the *sum total* of all well-being components. However, this is not the only possibility. One might prefer a function which takes the value of a life to be the *average* value of each well-being component. The function need not even be "aggregative" in this way. For example, the view that the value of $b$'s life is the value of the *worst* component, or the value of the *best* component, are also possible well-being functions.

The Same-World View takes the theory of well-being components and the well-being function as given. That is, it does not place any restriction on the well-being components themselves nor on the well-being function. The Same-World View is a claim about the contributive value of finally bad well-being components, not about what the finally bad well-being components are. According to this view, all finally bad well-being components make life go worse in the sense that a life is worse overall for a person taking the component into account. It is therefore a restriction on the *combination* of theories of well-being components and well-being functions.

One advantage of the Same-World View is that because there is no mentioning of possible worlds or similarity between possible worlds, we do not have to determine what would be the case if $S$ does not obtain. It is determinate what $b$'s life is like, and what well-being components it includes, and it is also determinate what $b$'s well-being is if we do not take a certain well-being component into account. The Same-World View is also more plausible than the Similarity View regarding overdetermination. In Back-Up, we can say that White's death is bad for her as long as our theory of well-being and the well-being function imply that her life is worse for her, taking her death into account, than it is not taking her death into account. So the fact that she

[15] The well-being function can favourably be compared with the *value*-function in Broome (2004, pp. 26–29). In Broome's case the value of a population is represented as a function of its members individual well-being while we are here letting an individual's well-being be a function of the good and bad things that befall her.

would have died in the most similar possible world (if that is indeed the case) is of no consequence for the contributive value of her death.

A further advantage of the Same-World View is that it is compatible with pluralism about the bearers of well-being. The Simple View and the Similarity View both assume that what has value for a person are propositions or states of affairs. On the Same-World View, however, things from different ontological categories can have final value for a person.

Note also that the Same-World View is a view about the contributive value of a state of affairs for a person and not the instrumental value of a state of affairs for a person. It is quite plausible that some counterfactual comparison relevant to the instrumental value of a state of affairs, and that the instrumental value of a state of affairs will for that reason depend on a counterfactual claim. The advantage of the Same-World View is rather that we can avoid the problem of overdetermination for contributive value. What is problematic about typical cases of overdetermination, like the one considered above, is whether an effect can be "attributed" to either Black's or Orange's act. Overdetermination should not be a problem for a theory about whether White's illness is bad for her, nor for a theory about whether White's illness makes her life worse.[16]

As I have formulated the Same-World View it seems to allow for an asymmetry between good and bad states of affairs with respect to their contributive value. On the Same-World View, it is not possible that a state of affairs which is bad for a person does not have negative contributive value. That is, all bad states of affairs makes a person's life go worse. It is possible, however, that a good state of affairs has negative contributive value. If it is false that *S* is bad for *b*, then nothing about *S*'s contributive value for *b* follows. The Same-World View is thus compatible with there being an asymmetry between good and bad states of affairs with respect to their contributive value.

However, this asymmetry seems to be ruled out if we consider that the Same-World View, if it is a plausible view about the contributive value of finally bad states of affairs, then it is just as plausible to analyse the contributive value of finally good states of affairs in an analogous way. It seems quite arbitrary to say that finally bad states of affairs make life go worse, but finally good states of affairs need not make life go better. We should then extend the Same-World View to good states of affairs as well: if a state of affairs is good for *b* then *b*'s life, taking the state of affairs into account, is *better* than *b*'s life not taking the state of affairs into account. A good state of affairs cannot, on this extended view, make life go worse and cannot be a harmful state.[17]

[16] In the next chapter I will consider the third condition in the basic structure, the responsibility condition, and argue that this should be understood as counterfactual dependence. A counterfactual comparison will therefore be relevant to the analysis of harm.

[17] It also seems plausible to extend the view to neutral states of affairs: if a state of affairs is neutral for *b* then *b*'s life, taking the state of affairs into account, is *equally as good as b*'s life not taking the state of affairs into account. This extension of the Same-World View rules out that neutral states of affairs can be harmful as well.

The Same-World View also allows for so-called "organic wholes". According to the principle of organic wholes, the intrinsic value of a whole need not equal the sum of its parts.[18] For example, a standard hedonistic theory which incorporates organic wholes could say that pain is the only thing which is intrinsically bad, but the whole of an episode of *deserved* pain is not intrinsically bad. On this imagined hedonistic view, if a person deserves pain then the pain taken by itself makes the person's life go worse. That is, the person's life is worse for her when we take this component into account. However, the Same-World View is compatible with the view that the complex component consisting of the pain and the fact that the person deserves to be in pain makes the person's life go better. Even though the two components are not independent of each other, the complex one cannot obtain without the simple one obtaining, they can nevertheless be evaluated independently. Taking the complex component into account does not force us to take the simple one into account.

The Same-World View sets up a close connection between the value of a state of affairs for a person and the contributive value of that state of affairs. For example, the view that pain is always bad for people, and that the value of a life is the average value of each well-being component, is not consistent with the Same-World View. An episode of pain can make a person's life go better, on this view, as long as it increases the average. According to the Same-World View, however, all bad states of affairs are such that they make a person's life go worse.

This casts some doubt over the claim that all states of affairs which are bad for a person makes that person's life go worse. As a *conceptual* claim about the relation between bad-for and contributory value the Same-World View seems mistaken since whether it is true depends on what the true theory of well-being is.

However, we need not reject the way contributory value is understood on the Same-World View because of this example. While it is perhaps a mistake to see the Same-World View as a claim about how good and bad-for relates to contributory value, as a claim about how we should understand the contributory value of a state of affairs the view still has merit. That is, the following claim is still plausible: if a state of affairs *S* has negative contributive value for *b* then *b*'s life, taking *S* into account, is worse than *b*'s life not taking *S* into account.

[18] The term organic whole was coined by Moore (1993) who claimed that "the value of [a whole] bears no regular proportion to the sum of the values of its parts" (Moore 1993, p. 79). It is a controversial issue exactly how such organic unities, or wholes, should be understood and whether they exist at all (see Carlson (1997)). Hurka (1998) distinguishes between a "holistic" view, which he claims was Moore's view on the matter, and a "conditional" view. On the holistic view the value of a part retains its value when it enters an organic whole and the additional value resides in the whole rather than in one of the whole's parts. On the conditional view, which Hurka ascribes to Korsgaard (1983), the value of a thing can change when it enters an organic whole.

Does this reinterpretation of the Same-World View undermine the arguments against the Similarity View, or the Simple View? No, it does not.

The main objection to the Simple View was that it is only plausible as a claim about final value, not about contributive value, because it does not take into account the difference a state of affairs makes to a particular life. This shortcoming of the Simple View is not remedied by switching to contributive value.

The main objection to the Similarity View was that it did not identify the contributive value of a state of affairs, *S*, but rather the contributory value of *S*-and-something-else. The Similarity View seems more plausible as a view about the overall value of a state of affairs, taking both final and instrumental value into account, but not about the contributory value of a particular state of affairs.

The Same-World View, I suggest, captures only the contributive value of a state of affairs, i.e., the sense in which a state of affairs makes life go worse. It might be suggested that when it comes to harm it is the contribution that a state of affairs makes and not its final value for a person which is relevant. That is, perhaps we should replace the second condition in the basic structure with:

> (2*): *b* is in a state *S* such that *b*'s life is worse, taking *S* into account, than *b*'s life not taking *S* into account.

This would leave it open whether *S* is finally bad and would instead emphasise the contribution of *S* to *b*'s well-being. One advantage of this suggestion is that we would not have to defend any particular relation between bad-for and worse-for. Now, a very persistent intuition regarding harm seems to be that to do harm is to make someone's life go worse (in some sense). This certainly favours (2*). Furthermore, consider the example with organic wholes above. If a bad state of affairs does not make a life go worse when we take it into account, because it is a part of an organic whole, then it seems inappropriate to say that the person has been harmed. Even though the state of affairs is bad for the person when considered in isolation, what we care about when it comes to harming seems to be the difference the state of affairs makes.

The difference between taking the contribution, rather than the value, of a state of affairs as relevant to harm need not be very great. It seems plausible that there is some correlation between the value of a state of affairs for a person and the state's contribution to the value of that person's life. This would, as we saw, rule out some combinations of well-being theories and well-being functions. However, one need not defend the Same-World View as a truth about the logic of prudential value. Instead, one could defend it as a substantial claim about well-being and contributive value. Recall that the Same-World View does not commit us to a simple, additive, well-being function. For example, it is consistent with the Same-World View that some bad states of affairs

make only a very small negative contribution, perhaps depending on when they obtain in a life or on what other states of affairs obtain in a life. One could therefore accept the principle of organic unities to a certain extent, for example. What the Same-World View commits us to is the claim that all bad states of affairs make *some* negative contribution. This is a much weaker and quite plausible claim.

As we saw above, it is possible on the Same-World View to endorse an asymmetry between good and bad states of affairs. That is, it is possible that some good states make a negative contribution to a life. However, this seems implausible. A more plausible view is that all good states make a positive contribution and all neutral states make no contribution to how well a life goes. With these assumptions we could make the stronger claim that a state of affairs *S* is bad for a person *if and only if b*'s life is worse for *b*, taking *S* into account, than not taking *S* into account.

There are then two independent routes by which we can arrive at (2*). First, by claiming that it is the contribution a state of affairs makes to the value of a life rather than the value of the state which matter to an analysis of harm. The second route is to claim that the value of a state of affairs for a person is the contribution which the state makes to the value of that person's life. The second condition in the basic structure, according to the second route, implies (2*).

### 4.3.1 The absence of benefits

I have argued that the way in which harms make life go worse is captured by the Same-World View and that we should replace the second condition in the basic structure with (2*). Compared to the other views discussed in this chapter (2*) seems to have a clear advantage. Unlike the Similarity View it avoids the problem of overdetermination and effectively isolates the contributive value of a particular well-being component. As we saw earlier, the Similarity View does not adequately capture the contributive value of a particular state of affairs, *S*, because the most similar world should plausibly be allowed to vary with respect to other states of affairs than *S*.

An advantage of the Similarity View, it might be argued, is that it gives a better account of the contributive value of absences of benefits. If we replace (2) with (2*) in the basic structure, then all harms necessarily involve a state of affairs which makes a person's life go worse. However, cases where a person is prevented from receiving a benefit, or where a person is deprived of a benefit, are often considered as examples of harm. Consider for example the following three cases:

> *Prevented benefits*: The quality of Red's life would be better were she to be employed but Magenta prevents this from happening.
> *Failure to benefit*: Magenta could improve Red's life by providing Red with employment but Magenta does not do this.

*Deprived of benefits*: Red has a full-time employment and her life is better for it. Magenta reduces Red's employment to a part-time position, lowering Red's quality of life. The part-time position still contributes positively to Red's life.

In each of these examples there is a case to be made for saying that Magenta harms Red, even though there does not seem to be a state of affairs which makes Red's life go worse. In each example, the ground for saying that Red has been harmed is that she could have been better off, not that there is something about her actual condition which contributes in a negative way to her well-being.

A possible reply to these examples is to agree that what Magenta does is in all three cases morally objectionable, other things being equal, but deny that this is because she harms Red. The tendency to say that Magenta harms Red when she does not provide Red with employment, for example, is because we think that people in general should benefit other people, other things being equal. The objection to Magenta's choice in this case is therefore not necessarily based on that Magenta harms Red but rather that Magenta should benefit Red. Likewise, we could object to Magenta's preventing Red from being employed on the same grounds. Magenta should not prevent Red from being employed because Red would benefit from being employed, not because Magenta would harm Red by preventing Red from being employed.

The intuition that Magenta harms Red in the examples above can therefore to an extent be explained away. Harm is sometimes used in a wider sense. It seems plausible to say that Magenta "harms" Red if one by this means merely that Magenta acts in an objectionable way towards Red, but that is "harm" in a wider sense than the one which I am interested in.

Not all cases of prevention, failure or deprivation of benefits can be explained away in this way. For example, a common view is that death is bad for a person only in so far as it prevents benefits. It would be a serious blow to the Same-World View if it rules out this account of the badness of death.

One way in which death can be bad for a person is because death is intrinsically bad for a person. This is not a very popular view, and perhaps for good reasons. One problem is that it cannot capture the plausible claim that the value of death for a person depends on when it occurs. For example, it is typically worse to die while young rather than while old, and in some cases it is perhaps not even bad to die. In cases where a life contains nothing but pain, for example, and there are no chances of relieving this pain in the future, then death might be neutral or even good for a person.

A more popular view about the value of death is the deprivation-approach.[19] According to the deprivation-approach, death is bad for a person because it prevents the person from enjoying future goods. In order for the Same-World View to be compatible with the deprivation-approach we therefore have to say that preventing a person from receiving a benefit is to harm that person.

[19] See for example Nagel (1991, ch. 1), Feldman (1991, 1992) and Bradley (2004).

In some cases, we would say that a lack of something good is bad for a person. A lack of food, shelter, fresh water, clean air, friends, education and so on are such cases. Such examples suggest that we should say that negative states of affairs can make a person's life go better or worse, not only positive ones. If such absences can make a person's life go better or worse, then it is possible to identify the state of affairs which makes Red's life go worse in the examples above, namely the state of affairs that Red does not have full-time employment. Likewise, the Same-World View would also be compatible with death being a case of deprivation. Suppose Magenta kills Red, thereby preventing Red from enjoying some future good *x*, so that the state of affairs "Red does not enjoy *x*" obtains. This state of affairs, I have suggested, does indeed make Red's life go worse because Red's life is worse for Red when this state of affairs is taken into account. It is also a state of affairs which Magenta's act is plausibly responsible for, though I will argue for that claim in the next chapter. We can therefore say that Magenta harms Red because the act of killing Red is responsible for a state of affairs which makes Red's life go worse, namely that "Red does not enjoy *x*".[20]

A natural objection is that it is implausibly wide to hold that all absent benefits have negative contributive value. For example, suppose Red's life is going very well. She has friends, her projects are successful, she is happy and so on. Her life could be better however. She could be even more successful, have even more valuable relations and be even happier. If the absences of a benefit makes a life go worse because negative states of affairs make a life go worse, then we would have to say that the state of affairs "Red is not more successful" makes Red's life go worse. But, it might be objected, it is not plausible that a person who is very well off is harmed just because she isn't even better off.

However, in so far as we find the original examples of absences of benefits plausible then we should not be moved by this objection. If we in the original examples think that it makes Red's life go worse that she misses out on possible good things, then I see no principled ground for saying that Red's lack of even more success and happiness does not make her life go worse.[21]

Furthermore, it does not follow that someone would harm Red just because there is a state of affairs which makes her life go worse. Unless someone is responsible for the state of affairs "Red is not more successful" then the only thing that follows is that this state of affairs makes Red's life go worse, not that anyone has harmed her. As far as doing harm goes, many negative states of af-

[20] It might be objected that this state of affairs can not make Red's life go worse because Red does not exist when it obtains. However, this is a general problem for the deprivation approach to death and not a problem which is unique to the Same-World View.

[21] Other views about the value of a state of affairs for a person have similar implications. For example, according to the Similarity Views discussed above it is bad for a person that she does not find Alladin's lamp because in the most similar world where she finds it her life, we can assume, goes better.

fairs will therefore be rather uninteresting because they obtain independently of what people do.[22]

To further strengthen this last point, suppose that someone were responsible for the fact that Red is not even more successful. Suppose that Magenta consistently sabotages Red's life so that she is not more successful. We then have a case which is very similar to our original three examples and in so far as we find these to be cases of doing harm then we should not object to saying that Magenta harms Red in this case.

In conclusion, absences of benefits do at least sometimes make a person's life go worse. Some examples involving the absence of benefits are not cases of doing harm. The tendency to think that there is harm done in such cases can be explained by appealing to related concepts such as moral permissibility and so forth. Other examples of the absence of benefits, in the case of absences caused by death for example, are indeed harms and the negative states of affairs which they refer to do make people's life go worse.

### 4.3.2 The value of existence

The view that we should include absent benefits among the things which make a person's life go better or worse raises the question what we should say about the value of existence and the value of a whole life for a person. If we say, as I think we should, that the state of affairs "Red is not more successful" makes Red's life go worse, should we then also say that the state of affairs "Red does not exist" makes her life go worse? On the face of it, the two states of affairs are similar, though it is far less intuitive to say that the latter has contributive value for Red. If Red does not exist, how can anything have value for her?

These puzzling problems also deserve special attention in relation to the Same-World View. Can the state of affairs "*b* exists" be bad for *b* on this view? Furthermore, what does the Same-World View imply regarding the kind of harms considered in the first chapter, i.e., states of affairs which seem to be bad for a person but where the person would not exist had the state of affairs not obtained?

Let us consider the latter question first. It should be obvious that the fact that a person would not have existed had a state of affairs not obtained is in no way relevant to the contributive value of the state of affairs according to the Same-World View. When we consider the the contributive value of a genetic condition like neurofibromatosis, or something even worse, the relevant comparison on the Same-World View is not with a possible state of affairs where the person in question does not exist. Likewise, whether having "a bad start in life" makes a person's life go worse is not a matter of considering how well

[22] Bradley (2004, pp. 60–2) argues in a similar way that it is not necessarily rational to have a negative attitude towards something just because it is bad for you. According to Bradley, it is bad for a person that she does not find Alladin's lamp but it is not something which anyone should care much about.

off the person would be in some different possible world. To have "a bad start in life" makes a life go worse if the life is worse, taking that state of affairs into account, than it is not taking that state of affairs into account. The fact that the existence of a person is contingent on the obtaining of a particular state of affairs is therefore no obstacle to the state of affairs being bad for that person.

Existence seems initially to be a more difficult matter. A common view is that the state of affairs "*b* exists" cannot be good or bad for *b* because that would require "*b* exists" to be better or worse than the state of affairs "*b* does not exist". However, if "*b* exists" is, say, better for *b* than "*b* does not exist" then it seems to follow that "*b* does not exist" is *worse* for *b* than "*b* exists."[23] But, it seems counter-intuitive to say that a non-existent person who would have a life worth living, were she to exist, is worse off not existing. For example, my merely possible third uncle, who does not in fact exist, is neither better or worse off than he would be had he existed. Nothing has value for my third uncle simply because he does not exist.[24]

On the other hand there is a case to be made for saying that existence can have value for a person. It does at least seem as if we evaluate the value of existence when we say that a life is "(not) worth living" or that someone is "better off dead" for example. A way to accommodate such intuitions is to claim that non-existence is neutral for a person, and that existence is good for a person when it is better than non-existence.[25]

According to the Same-World View, both of these views rest on a mistake in so far as they intend to address the question whether *existence* can have contributive value for a person. Whether existence makes a life better or worse, on the Same-World View, is a matter of whether a life is better taking existence into account than it is not taking it into account. That is, the value of existence does not depend on a comparison with non-existence. If, however, they are merely claims about the comparative value of existence as compared with non-existence, then they do not address whether existence as such can have contributive value for a person.

On the Same-World View, whether existence has contributive value for a person, *b*, depends on if *b*'s life, taking the state of affairs "*b* exists" into account, is better or worse than the value of *b*'s life not taking this state of affairs into account. Recall that the Same-World View does not place any restriction on what a substantial theory of well-being should look like. To take the state of affairs "*b* exists" into account is merely to treat it as a well-being component, and to not take it into account is to not treat it as a well-being component. The

[23] See McMahan (1988), Broome (2004) and Arrhenius (2013).

[24] Several authors have suggested that existence can be better (or worse) for *b* in a world where *b* exists while it also being the case that non-existence *would* not be worse (better) for *b* (see Arrhenius & Rabinowicz (2010) and Johansson (2010)). This is controversial however because we would then have to reject the "accessibility principle", argued for by Bykvist (2007*a*), that if *S* is better (worse) than *S*′ for *b*, then *S* would be better (worse) even if it obtained.

[25] See for example Holtug (2001) and Roberts (2011).

Same-World View does therefore not require us to make any dubious comparison with how well-off *b* would be if *b* did not exist.

There is a further difference between the Same-World View and the two views above which should be emphasised. It is commonly assumed that if existence is good for a person then it follows that non-existence is bad for a person. This is so because, as we saw above, that existence is good is understood in terms of a comparison with non-existence. However, on the Same-World View this is not the case. On the Same-World View the value of the state of affairs "*b* exists" depends on the difference this state of affairs makes when we take it into account. It does not depend on whether it is better for *b* to exist rather than not. The value of existence and non-existence should therefore be considered separately.

Let us consider non-existence first. It seems very intuitive, given the Same-World View, to say that non-existence is neutral for a person. Taking the fact that a person does not exist into account does not seem to be better for the person than if we do not take non-existence into account. A reason for making this claim is that we are making the comparison "within" the same world, so to speak. Whether we take the person's non-existence into account or not does not change the fact that the person does not exist, and nothing has value for a person who does not exist simply because there is no one for which the thing can have value. Of course, *if the person were to exist* then many things would have value for this person. We can formulate this by saying that in those possible worlds where this person exists, things have value for her. However, in the possible worlds where she does not exist, nothing has value for her.

This suggests that while non-existence is not a well-being component in its own right it can still make a difference to a person's well-being in that it can *undermine* the value of other states of affairs. A state of affairs which has value for a person in worlds where she exists has neutral value for this person in worlds where she does not exist. A person's well-being therefore depends on the person's non-existence but it does not follow from this alone that non-existence has value for this person. We can distinguish between what a person's well-being depends on and what constitutes a person's well-being. It seems plausible that well-being can depend on things which does not constitute it. For example, on a hedonistic view a person's well-being depends on the fact that she has consciousness. Otherwise, she would not be able to experience pleasure or pain in the first place. However, consciousness in itself does not constitute a person's well-being according to hedonism, only specific types of experiences do.

Now consider the value of existence. Considering the reason for saying that non-existence has neutral value for a person it seems plausible to say that existence also has neutral value. When we take the fact that a person exists into account we do not judge that person as better off than if we do not take it into account. The reason is again that we are making comparisons within the same world and not changing the fact whether the person exists. Existence

does seem to make a difference, however, in the same way as non-existence makes a difference. The fact that a person exists in a possible world *enables* other states of affairs to have value for this person but it is not, in itself, good or bad.

What about the plausible claims mentioned above, that a life can be worth living and that some people can be better off dead? On the view suggested here, such claims can be understood as being about the distribution of finally good and bad things in a person's life rather than about the final value of a person's existence. When we say that *b* has a life worth living what we mean is not that *b*'s *mere* existence is good for *b*. Also, we need not claim, according to the view suggested here, that non-existence would be better for *b*. What we should say is that the good things in *b*'s life outweigh the bad things. This, it seems to me, is the most straightforward and plausible account of such claims.

Finally, on the view just sketched, we can claim that existence and non-existence are neutral for a person without having to say that non-existent people, such as my third uncle, are well-off to some extent. Because existence is a precondition for other things having value for my third uncle, and this condition is not satisfied in his case, nothing (not even that he does not exist) has value for him.

## 4.4 Summary

In this chapter I have considered the claim that a necessary condition for doing harm is that there is a state of affairs which is finally bad for someone. However, to do harm is also to make someone's life go worse, in some sense. The aim of this chapter has been to spell out the latter: in what sense does bad states of affairs make a person's life go worse? The view I favour is the Same-World View:

> *The Same-World View*: if a state of affairs *S* is bad for *b* then *b*'s life, taking *S* into account, is worse than *b*'s life not taking *S* into account.

I have argued that it is possible to strengthen the Same-World View given that we treat good and neutral states of affairs in the same way. If we do, then the value of a state of affairs *is* the difference it makes to the value of a life. That is, we could replace (2) in the basic structure with

> (2*): *b* is in a state *S* such that *b*'s life is worse, taking *S* into account, than *b*'s life not taking *S* into account.

A consequence of the Same-World View, I argued, is that a state of affairs can make that person's life go worse even though the person would not have existed had the state of affairs not obtained. This is because on the Same-World View the relevant comparison is "within" a given possible world and not with

some other possible world. This is relevant to the non-identity problem because it shows that the fact that the young girl's child would not exist had the girl postponed her pregnancy simply is not relevant to whether the child's bad start in life makes her life go worse. Likewise, whether neurofibromatosis is makes the Specks' third child's life go worse is in no way dependent on the fact that that child would not have existed had either of the two physicians not been negligent. What *is* relevant is whether the children in question have lower well-being when we take the bad start in life, or the neurological condition, into account.

# 5. Responsibility

In the previous chapter I argued for a view about the contributive value of harm, the Same-World View, and that we should replace the second condition in the basic structure with (2*). Keeping to the basic structure, we then have the following partial analysis of harm:

> *a* harms *b* only if
> (1) *a* performs an act, $\phi$,
> (2*): *b* is in a state *S* such that *b*'s life is worse, taking *S* into account, than *b*'s life not taking *S* into account.
> (3) $\phi$ is responsible for *S*'s obtaining.

In chapter two I suggested that (3) seems to be a necessary condition for doing harm. For a person do harm it is not enough that the person performs an act and that there is a harmful state of affairs. The harmful state of affairs, I claimed, must also be related in the right way to the act. That is, the act must be responsible for the effect.

In this chapter I will argue that we should analyse the third condition in terms of counterfactual dependence. More precisely, I will argue that the third condition should be understood in the following way:

> (3*) if $\phi$ had not been performed then *S* would not have obtained at all, or would have obtained in a different way which would be salient in the circumstances.

According to this view, an act is responsible for a state of affairs just in case the act makes a difference to whether the state obtains at all, or whether it obtains in the same way. This captures nicely a central aspect of how we think about the consequences of acts. If an act makes no difference whatsoever to the obtaining of a state of affairs then this strongly suggests that the act is not responsible for the state of affairs.

It is important to distinguish the kind of responsibility which I will be concerned with in this chapter from responsibility in a broader sense. As I have already indicated, "responsibility" as I will use the term is a relation between an act and an effect such that the effect is a consequence of the act, or alternatively, that the effect can be "attributed" to the act.[1] Responsibility is

[1] A terminological note: it is a debated issue whether "the effects" of an act are events, states of affairs, or something else. In what follows I will use "effects" and "states of affairs" as

sometimes used in a different sense. One sense of responsibility which differs from the one I am interested in is when we ascribe responsibility to a person merely to indicate that the person is under a special obligation. Parents, for example, may have this kind of responsibility for their children. Responsibility in this sense is clearly normative because it involves a claim about someone's duties. The sense of responsibility which I am interested in is in itself non-moral though it might have moral relevance. Saying that $\phi$ is responsible for *S*'s obtaining does not, as I will use the term, involve any claims about duties, praise or blame. A further difference is that responsibility in this latter sense is what is typically indicated when a person "takes" responsibility for something or someone. This is not the sense of responsibility which I will be concerned with in this chapter. Whether an act is responsible for an effect is not something which the agent can make happen by "taking" responsibility.[2]

Though *prima facie* intuitive, analysing responsibility in terms of counterfactual dependence faces some well-known problems. Most notably, such analyses gives the wrong result in two familiar cases: Overdetermination and Pre-Emption. I have already discussed these cases to some extent in earlier chapters. There I argued that examples of this kind provide a good reason for rejecting the Counterfactual Condition as a necessary condition for doing harm. In this chapter, I will argue that the failure of the Counterfactual Condition does not force us to reject analysing responsibility in terms of counterfactual dependence. That is, understanding the responsibility condition in terms of counterfactual dependence does not entail the Counterfactual Condition and the most plausible defence of counterfactual dependence against Overdetermination and Pre-Emption does not lend any support to the Counterfactual Condition.

Finally I will consider the claim that there is a simpler and more straightforward notion of responsibility which avoids the problems that plague counterfactual dependence. A natural suggestion is that an act is responsible for a state of affairs simply by *causing* that state of affairs to obtain. I will argue that the advantages of this "causal view" are merely apparent and that there are significant differences between causation and responsibility which provide sufficient reason to reject the causal view.

---

interchangeable. This should not be seen as a heavy ontological commitment however. The main reason for choosing states of affairs is that (2*) is formulated in terms of states. But, note that one feature of (2*) is that it is compatible with other things than states of affairs having negative contributory value. The choice of states of affairs as the favoured ontological category is therefore one of convenience, not of necessity.

[2] See Scanlon (1998, pp. 248–9). Scanlon distinguishes between "responsibility as attributability" and "substantive responsibility". The former corresponds to the sense of responsibility which I am concerned with in this chapter and the latter corresponds to the broader notion of responsibility.

## 5.1 Counterfactual dependence

A common view about responsibility is that an act is responsible for a state of affairs just when the act makes a difference to that state's obtaining. If a state of affairs would have obtained independently of an act then this strongly suggests that the act is not responsible for the state's obtaining. Taking this plausible claim as our starting point we can formulate the following counterfactual analysis of responsibility:

> *The Difference-Making View-1*: an act, $\phi$, is responsible for a state of affairs, $S$, if and only if, if $\phi$ had not been performed then $S$ would not have obtained.[3]

The right-hand side of the Difference-Making View is usually understood in terms of possible worlds, and I will follow suit in this respect. That is, $S$ depends counterfactually on $\phi$ if and only if the closest $\neg\phi$ world to the actual world is a world where $S$ does not obtain. For example, if I strike a match ($\phi$) and there is an explosion ($S$) then my striking of the match is responsible for the explosion if and only if there is no explosion ($\neg S$) in the most similar world where I do not strike the match ($\neg\phi$).

A well-known problem for the Difference-Making View is that it yields the wrong result in two cases which are familiar from chapter two:

> *Pre-Emption*. Black poisons White. Before the poison has any effect Orange kills White. Had Orange not killed White, then the poison would have killed White just a moment later.
> *Overdetermination*. Black and Orange, independently of each other and at the exact same time, shoot White. Each shot is sufficient to kill White.

In chapter two it was claimed that a satisfactory solution to these two cases should imply that someone harms White in Overdetermination and that at least Orange, though perhaps not Black, harms White in Pre-Emption. A theory of responsibility should give similar verdicts in these cases. It is clear that someone is responsible for White's death in Overdetermination and that at least Orange, though perhaps not Black, is responsible for White's death in Pre-Emption.[4]

The Difference-Making View does not satisfy these two criteria because in both cases, White's death is not counterfactually dependent on either Black's

[3] The Difference-Making View has obvious similarities with counterfactual analyses of causation. See Lewis (1973*a*, 1979, 2000). The relation between a counterfactual analysis of responsibility and causation will be discussed further below.

[4] From a pre-theoretical point of view, would we say that Black is responsible for White's death in Pre-Emption? This seems far from obvious. After all, the effects of Black's act never occur so it would seem odd to hold Black responsible for White's death. However, Black's act seems to be morally objectionable because of what it would lead to under "normal" circumstances. I will pass over this controversy. For our purposes, we only need the weaker claim that Orange is responsible for White's death in order for Pre-Emption serve as a counterexample to the Difference-Making View.

or Orange's acts. In Pre-Emption, White's death is counterfactually dependent on Orange's act only if White's death does not occur in the closest possible world where Orange does not kill White. However, if Orange and Black act independently of each other, then the world where Orange does not perform his act is one where Black still performs his act. So White would still die, though slightly later, in the closest possible world where Orange does not kill White.[5] Since the same reasoning can be applied to Black's act, it follows that neither Black's nor Orange's act is responsible for White's death. The same reasoning also leads to the conclusion that neither Black's nor Orange's acts are responsible for White's death in Overdetermination. However, it is clear that at least Orange's act is responsible for White's death in Pre-Emption and that either Black's or Orange's acts, or both, are responsible for White's death in Overdetermination.

A general approach to both Pre-Emption and Overdetermination starts with the observation that Pre-Emption and Overdetermination are well known problems for theories of causation as well. A plausible *sufficient* condition for "*c* causes *e*" is that *e* depends counterfactually on *c*: had *c* not obtained then *e* would not have obtained. Pre-Emption and Overdetermination, often summed up under the heading of "redundant causation", serve as serious counterexamples to the claim that counterfactual dependence is necessary as well as sufficient for causation because it implies that Orange does not cause White's death in Pre-Emption and that neither Black nor Orange cause White's death in Overdetermination.[6] A way to approach the problem would then be to consider the responses to such redundant causation and see if they can be used to formulate a view about responsibility in terms of counterfactual dependence.

### 5.1.1 Redundant causation

There are, broadly speaking, two ways to approach the problem of redundant causation which are relevant here: an "individual" and a "collective" approach. In terms of causation, the individual approach is that Black and Orange cause White's death on their own in Overdetermination while on the collective approach they cause White's death together.[7]

I have discussed the collective approach in chapter two and will only briefly restate why it is not a plausible view. For the collective approach to be plausible, it has to be clarified in what sense Black and Orange act "together". One suggestion is that Black and Orange act together in the sense that they

[5] The notion of the "closest" possible world is of course a bit unclear. I assume that the distance between two worlds depends on their similarity, and that similarity (as I argued in the previous chapter) should be understood in such a way that the most similar world where Black does not act is one where Orange still acts, provided that they act independently.

[6] See for example Lewis (1973*a*, 2000).

[7] For an overview of the debate and a defence of the individual approach, see Schaffer (2003).

perform a collective action. This collective act, it is then claimed, is responsible for White's death because White's death depends counterfactually on the performance of the collective act. However, this version of the collective approach is not plausible because it would stretch the notion of "collective act" in unacceptable ways. There is a difference between Black and Orange together doing something and their acts contributing to the same outcome but the "collective-act" reply to Overdetermination and Pre-Emption does away with this distinction.

A further problem for this version of the collective approach is that it presupposes that if a collective act is not performed then none of the individual acts which constitute the collective act is performed. But this is not the case. It is sufficient for the collective act "you and me painting the house" not to be performed that I fail to do my part. You can paint as much as you like and the collective act will still not be performed. This is especially troubling in Overdetermination where it then seems that the collective act "Black and Orange fire their guns" is not performed but that Black's act is.

Alternatively, the collective approach could be understood as saying that Black and Orange are both responsible for White's death in overdetermination because White's death depends counterfactually on their acts taken together. Black's and Orange's acts are members of a minimal set such that White would not have died had none of the acts in this set been performed.

A problem for this version of the collective view are cases where there are more than one minimal set which the effect depends counterfactually on. For example, consider (again) the following case:

> *The Death Squads*. Brown is choosing which death squad to join, *A* or *B*. These two groups will then attempt to catch members of The Resistance and execute them. Brown knows that as long as a death squad has at least ten members it will be successful and that group *A* will catch and execute 1000 people while group *B* will catch and execute 10 people. If a squad has fewer than ten members then it will fail to catch any members of The Resistance. Brown also knows that group *A* has ten members and group *B* has nine.

Suppose that Brown is only concerned with minimising the amount of harm he would do. If he joins *A* then the set consisting of him and any one of the other members of *A* form a minimal set such that the 1000 deaths would not have obtained had none of this minimal set's members joined *A*. So he would be responsible for 1000 deaths were he to join *A*. If he were to join *B*, on the other hand, then he would only be responsible for ten deaths. This leads to the absurd conclusion that Brown should join *B* if he wants to minimise the harm he does.

In chapter two I also argued that attempts to save the collective view by saying that Brown would do less harm, but still some harm, if he were to join group *A* fails. Even if the Brown's responsibility for the 1000 deaths depends

on the size of the group he joins it is still the case that Brown would do more harm by joining *B*.

We should therefore consider the individual approach and see whether it has the resources to yield more plausible answers to Pre-Emption, Overdetermination and the Death Squads.

It might be claimed that the problem with redundant causation arises because we have, when framing the examples, presupposed that there is only one effect involved, namely "White's death". However, this is to presuppose a "coarse" individuation of effects where differences to when, where and how an effect occurs does not affect the identity of the effect. The solution is to adopt a view of effects where the identity of an effect is fragile. On this view, very small differences with respect to when, where or how an effect occurs affect whether it is the same effect. For example, we would not say that "my death" denotes the same death whenever it occurs. I am bound to die some day, but dying in a car accident tomorrow and dying of old age in 50 years are clearly different deaths. Applied to Overdetermination, one could on this view say that Black's and Orange's acts have different events as their effects: "death by Black" and "death by Orange".[8]

It is questionable however whether this is the correct way to describe the case. Saying that Black's and Orange's acts have different effects does not sit well with the description of the case as a situation where their acts are independently sufficient for an effect. In Overdetermination especially there seems to be just one effect: White's death. While it is plausible that some changes as to when, where and how an effect occurs can affect the identity of the effect, it does not seem plausible that *any* such change will affect the identity of the effect. In Overdetermination, for example, the difference between "death by Black" and "death by Orange" is too insignificant to make them into different effects. A further problem for this approach is that it identifies seemingly irrelevant factors as relevant to event-identity. On the fine-grained approach White's actual death would be counterfactually dependent on many factors of the environment. For example, had the wind been blowing differently then Black's (and Orange's) bullets would have hit White in a slightly different way. According to the Difference-Making View it would then follow that the way the wind blows is also responsible for White's death. But we would not say that the exact way the wind blows is responsible for White's death, nor that it is relevant to the identity of White's death.[9]

A more plausible approach to redundant causation is to adopt a coarse individuation of effects but to distinguish between the different ways in

[8] Note that it does not follow, on this approach, that Black and Orange make a difference to White's well-being since it is not worse for White if both "death by Black" *and* "death by Orange" obtain. See also below where I compare the Counterfactual Condition with the Difference-Making View.

[9] This problem for the fine-grained version of the individual view is raised by Lewis (2000) and Petersson (2004).

which White's death might obtain.[10] Within theories of causality, the idea is that Black causes White's death in Overdetermination because White's death would not have occurred at all *or* would not have occurred in the same way had Black not acted as she did. The counterfactual approach to responsibility could then be modified in the same way:

> *The Difference-Making View-2*: an act, $\phi$, is responsible for a state of affairs, *S*, if and only if, if $\phi$ had not been performed then *S* would not have obtained, *or*, *S* would not have obtained in the same way.

On this view effects are not fragile and the same effect can occur in different ways. The condition is also weaker than the original counterfactual analysis where *c* causes *e* if and only if *e* would not have occurred had *c* not occurred. On the current view it is enough that *e* does not occur in the same way, had *c* not occurred.

The Difference-Making View-2 is rather rough, mainly because the notion of the "ways" in which an effect can occur is in need of clarification. The most obvious problem is to characterise the identity-changing ways of an effect. In Overdetermination, for example, there seems to be only one effect involved: White's death. Had White only been hit by Orange's bullet then she would have died the same death. The difference in the way which White's death occurs does not make her death into a different effect according to this new version of the Difference-Making View. But, this certainly calls out for an explanation. However, for our present purposes we can safely ignore this problem. The Difference-Making View-2 only requires that the effect *either* does not occur at all, had the act not been performed, *or* that the effect does not occur in the same way, and this disjunction is true of both Black's and Orange's acts regardless of how the notion of "ways" is further clarified. I will therefore leave it open exactly how to draw the distinction between identity-changing and identity-preserving features of an effect.

A more serious problem for this view is to identify what the ways of an effect are to begin with. Consider again White's death. In order for the Difference-Making View-2 to solve the problem of Overdetermination we have to assume that dying from two bullets rather than one is either a different death entirely or a way in which White's death could occur. But, there are further features of this particular death where it is perhaps less clear whether they are ways of White's death at all. For example, suppose that White dies in the shade. Should we then say that dying in the shade is a way in which White's death occurs? If it is, then the individualistic approach amounts to a much broader notion of responsibility than we commonly acknowledge. Any act which makes a slight difference to the circumstances in which White's death occurs would, according to the Difference-Making View-2, be responsible for White's death.

[10] See Paul (1998, 2000), Lewis (2000) and Schaffer (2003).

Broadening the notion of responsibility in this way threatens to make the individual approach implausibly wide. It would also make the Difference-Making View-2 harder to distinguish from the collective approach in a case like the Death Squads. If Brown makes a difference to the way in which the members of The Resistance would die by joining the squad with ten members then he would be responsible for their deaths according to this version of the Difference-Making View.

A plausible suggestion at this point is that we could avoid these problems for the individual approach if we could make a distinction between features of White's death on the one hand and features of the situation on the other. Dying in the shade is intuitively not a way in which White's death occurs but rather a feature of the circumstances more broadly construed. One way to draw this distinction is by saying that dying in the shade is not an intrinsic property of White's death and that it is therefore not a "way" in which White's death occurs. However, maintaining this distinction on such metaphysical grounds seems to be difficult. For example, the time when White's death occurs is perhaps not an intrinsic property of White's death but it should be considered a way in which it occurs. Otherwise the Difference-Making View-2 would not be able to handle cases of Pre-Emption.[11] I will therefore not pursue this option here.

Instead, I suggest that we adopt a more conventional approach. The difference between features of the circumstances and features of White's death are, I suggest, to be explained by their relative salience.

The view that it is conventional matters which distinguish causal conditions and primary causes has a long tradition.[12] One example is Mackie (1955) who argues that a conventional approach is required to capture ordinary talk about causation:

> in determining responsibility we do not choose the causal field [i.e., the set of conditions which are jointly sufficient for an effect] quite arbitrarily; our choice is determined by our moral expectations, our views about what is normal and proper (Mackie 1955, p. 145).

This view is also present in Mackie (1980) where he, in opposition to the view of Hart & Honoré (1985), argues that "there is a single basic concept of causing to which various frills are added" (Mackie 1980, p. 117). This basic concept is of course his analysis of causation in terms of INUS-conditions. The "frills" which can be added to this core are necessary in order to distinguish between primary causes and causal conditions for example. More recently, Paul (1998) and Lewis (2000) both suggest similar views where causa-

[11] See also Paul (1998) for further arguments against excluding so-called "hasteners".

[12] The view is often attributed to Mill who in *A System of Logic* (Bk. III, ch. 5, sect. 3) finds no real distinction between causes and conditions. I will not enter into any historical debate regarding how far back this view goes, though I think there is reason to go at least as far back as Hume. See Mackie (1980, chs. 1–2).

tion is analysed in terms of counterfactual dependence but where conventions and salience can play a role to explain why we tend to treat some features of White's death as relevant and some not.

Among these views it is common to identify the primary cause of an effect with the most salient cause in the circumstances. For example, if I light a match and there is an explosion, then whether I caused the explosion depends on whether my act was salient in the circumstances. If I struck the match while at a gas-station then my act would be the primary cause since in such circumstances the presence of inflammable gasses are quite normal. However, if I struck the match while at home, and there was an explosion, then it would be the presence of inflammable gasses rather than my act which would be salient.

Note however that using salience in this way to distinguish between causes and conditions does not help us explain why dying in the sun is not a way in which White dies. In fact, the cause of "White dies in the sun" might be very salient in the circumstances.[13] If salience is to help us distinguish between the ways of an effect such as White's death on the one hand and features of the situation on the other then it is the salience of these features, not the salience of their causes, which is relevant.

We should therefore expect the relative salience of a feature to depend on a number of factors, many which are conventional in nature. One such factor is the one pointed out by Mackie in the quote above. Whether dying in the shade is a way in which a particular death occurs or whether it is a feature of the circumstances depends on whether dying in the shade is perceived as normal or abnormal in the circumstances. Dying in the shade is under most circumstances not a very salient feature because dying in the shade is typically no more abnormal than dying in the sun.[14]

A further factor which contributes to the salience of a feature is the difference this feature makes to the effect in question. For example, dying a year later than one would otherwise have done is more salient than dying a second later because the former makes a greater difference to the death than the latter.

The salience of a feature can also depend on general causal correlations. If there is a causal correlation between the way in which an effect occurs and the effect itself, then this correlation contributes to the salience of the way in question. For example, there is for normal human beings no general causal correlation between being in the sun and dying. However, there is a general causal correlation between being hit by a bullet and dying from a bullet.[15]

---

[13] Suppose White dies at night, but someone has arranged a giant mirror in space which reflects sunlight to White.

[14] If I am allowed a fantastic example, a case where dying in the sun is not a normal feature of a death is when the victim is a vampire.

[15] General causal correlations are of course compatible with there being exceptions when it comes to particular causal relations. For instance, it is a general causal correlation that smok-

This lack of correlation makes dying in the sun a less salient feature than dying from a bullet.

With the notion of salience we can formulate a third version of the Difference-Making View:

> *The Difference-Making View-3*: an act, $\phi$, is responsible for a state of affairs, *S*, if and only if, if $\phi$ had not been performed then *S* would not have obtained at all, or would have obtained in a different way which would be salient in the circumstances.

With this view one can argue that Black is not responsible for White's death in Pre-Emption in the following way. Black's poisoning does not make a difference to whether White's death occurs, and it makes no difference to the way in which White's death occurs, so he cannot be responsible. Orange, on the other hand, is responsible for White's death because making White's death occur earlier is a salient feature in the circumstances. This verdict, as I claimed above, is acceptable.

In Overdetermination we can in a similar way claim that both Black and Orange make a salient difference to the way in which White's death occurs. Had Black not fired his gun then White would have died only from one bullet rather than two, and the number of bullets which hit White is salient in these circumstances.

It could be objected however that if, as I just suggested, Black and Orange are responsible for White's death in Overdetermination then we cannot at the same time say that Brown would not be responsible for the 1000 deaths if he joins group *A* in the Death Squads. If the difference Black and Orange make in Overdetermination is salient, then the difference Brown makes in the Death Squads should also be salient. This would make the Difference-Making View-3 more similar to the collective version which I rejected above.

Suppose we grant that the two examples, the Death Squads and Overdetermination, are sufficiently similar and that Brown would do harm if he were to join group *A* in the Death Squads. The individual approach still has a distinct advantage over the collective approach here because we can say that there is a difference in the degree of responsibility between joining group *A* and group *B* in the Death Squads. This becomes clear if we consider the salience of the difference Brown's choice makes. If he joins group *A* then the difference he makes will not be very salient compared to the difference he would make were he to join *B*. If he were to join *B* then he would make a difference to whether the effect occurs at all, but if he joins *A* then he only makes a difference to the way in which 1000 deaths occur. We can therefore claim that Brown would do less harm by joining *A* because the difference he makes is smaller.

---

ing causes cancer. However, this general correlation can hold even if particular instances of smoking does not lead to cancer.

Note that I am not suggesting that Brown would only be responsible for the difference he makes in these cases. Applied to Overdetermination that view would have the unintuitive consequence of making neither Black nor Orange responsible for White's death, strictly speaking. What I am suggesting is that responsibility is not an all or nothing matter but rather a matter of degrees. The claim is that Brown would be less responsible for the outcome, 1000 deaths, if he joins *A* than he would be for the outcome, ten deaths, if he joins *B* and that the degree of harm he does should be adjusted accordingly.[16]

An explanation of this kind is not available to the collective approach. On the collective approach there is no way of distinguishing between different degrees of responsibility based on different degrees of contribution. If Brown is responsible for 1000 deaths if he joins *A* because of his membership in a set which has certain properties, then there is no ground for holding him more or less responsible for that outcome than other members of this set, even if there is a clear intuitive difference in the degree of individual contribution to the outcome. The individual approach therefore has a clear advantage over the collective approach in this respect, despite having similar consequences in a case like the Death Squads.[17]

The Difference-Making View-3 captures the plausible claim that responsibility is a matter of making a difference. It is also more plausible than the collective approach regarding Overdetermination, Pre-Emption and the Death Squads. Assuming this view, the third condition in the basic structure can therefore be replaced with:

> (3*) if $\phi$ had not been performed then *S* would not have obtained at all, or would have obtained in a different way which would be salient in the circumstances.

This analysis of responsibility is similar in some ways to the Counterfactual Condition from chapters one and two. Before proceeding to discuss an alternative analysis of responsibility it is therefore necessary to consider the relation between the Difference-Making View and the Counterfactual Condition.

---

[16] See Roemer (1993) for an alternative theory of responsibility which allows for degrees. However, Roemer's notion of responsibility differs slightly from mine and it is therefore unclear to what extent my view conflicts with his.

[17] See Schaffer (2003) for a similar argument. According to Schaffer, the collective approach to Overdetermination fails because it "cannot offer a stable account of the causal contribution of individual overdeterminers" (Schaffer 2003, p. 38). The problem, according to Schaffer, is that the collective approach should grant that individual overdeterminers cause *something*. Otherwise the approach would be, he claims, "vastly implausible". However, with such an account of the causal powers of individual overdeterminers the collective approach turns out to be just a version of individualism.

### 5.1.2 Counterfactual dependence and the Counterfactual Condition

In chapter two I argued that we should reject the following condition:

> *The Weak Counterfactual Condition*: An act harms a person only if that person is worse off than she would have been had the act not been performed.

The main reason was that the Weak Counterfactual Condition rules out that Black and Orange harm White in Overdetermination, and that Orange harms White in Pre-Emption. But, if the Weak Counterfactual Condition can be defended on similar grounds as those use to defend the Difference-Making View then we would have to reconsider the rejection of the Weak Counterfactual Condition.

First, it should be clear that the Weak Counterfactual Condition and the Difference-Making View are two distinct views about the necessary conditions for doing harm. Different-number cases can be used to illustrate the difference. Consider the case with the young girl. Suppose the girl does not wait, that is, she has a child while young and the child gets a "bad start" in life. The Counterfactual Condition, on any version which I considered in chapter two, is clearly not satisfied in this case. The child is not worse off than she would otherwise have been because her life is, despite the bad start, worth living. However, it is equally clear that the child's bad start in life is counterfactually dependent on the girl's choice: if the girl were to postpone her pregnancy then the bad start in life would not obtain. The Difference-Making view does therefore not imply the Weak Counterfactual Condition.[18]

Turning to the question whether the Weak Counterfactual Condition can be defended on similar grounds as I have defended the Difference-Making View the answer must be "no". It is plausible given the Difference-Making View that Black and Orange are responsible for White's death in Overdetermination because the way in which White's death occurs depends on their acts. But, this dependence does not help the Counterfactual Condition. Even if it is true that White's death would not have occurred in the same way, had not Black acted as she did, it is still the case that White's death would have occurred. It is therefore still false that White would have been better off had Black (or Orange) not acted as he did. Even though White's death would have occurred in a different way had Black not acted as he did, this difference would not make White any better off. The difference in the way in which White's death occurs does not mark a difference in the value of White's death for White. This shows that the individual approach is not available to the Weak Counterfactual Condition as a solution to Pre-Emption and Overdetermination.

In very general terms, the problem for the Weak Counterfactual Condition with respect to Overdetermination can be summed up as follows. Suppose

[18] This point is also made by Hanser (1990).

that we have a theory of responsibility which solves the problem of Overdetermination and Pre-Emption. That is, the theory implies, among other things, that Black and Orange are responsible for White's death in Overdetermination. However, this theory of responsibility does not change the fact that it is still *false* that White is worse off than she would have been had Black not done what he did. Solving the problem of Overdetermination for the Counterfactual Condition is therefore distinct from solving the same problem for responsibility, and successfully doing the latter will not, *ipso facto*, lead to successfully doing the former.

## 5.2 Responsibility and causality

Responsibility, I have suggested, should be understood in terms of counterfactual dependence. In doing so I have drawn on refinements to the simple counterfactual analysis of causation and adapted those to an analysis of responsibility, for example, the idea that an act can be responsible for an effect if it makes a difference to the way a state of affairs would obtain. I also claimed that general causal correlations are relevant to the notion of salience which played a central part in distinguishing the individual from the collective approach. It might be wondered however whether this analysis of responsibility is in fact parasitic on the analysis of causation in terms of counterfactual dependence. That is, perhaps a less committing though just as plausible view would be to say that an act $\phi$ is responsible for a state of affairs $S$ if and only if $\phi$ causes $S$ to obtain. On this view, we can leave questions of causality to metaphysics and go with an intuitive, common-sense understanding of causality. Call this view *the Causal View*.[19]

The main advantages with the Causal View are methodological. First, it is a simple view. Leaving causality unanalysed allows for less complicated analyses of harm. Second, on the Causal View we can, at least in the present context, ignore problematic cases like Pre-Emption and Overdetermination. This would allow us to focus on what is really problematic about, for example, same-number cases like the young girls' choice. What is problematic about the Non-Identity Problem is not whether we are responsible for harmful effects in the future but whether one can do harm to a person if one could not have made that person better off. As the case of the young girl illustrates, everyone in the

[19] See Hart & Honoré (1985). They argue that questions of responsibility in the law are questions of causality, and criticise the "minimalists" who "allot only a minor role to causal issues in determining questions of legal responsibility, and who for the most part hold that the only genuine causal issue is that of *sine qua non*" (Hart & Honoré 1985, p. lxvii). The Causal View also seems to be assumed by Mackie (1955). More recently the Causal View has been endorsed (though not explicitly defended) by Harman (2009). Braham & van Hees (2012, p. 605) suggests a more modest view where causation is necessary for responsibility. As I will argue shortly, we should not even accept this more modest claim.

debate can agree that she causes her child to have "a bad start in life" but this does not settle the question whether she harms her child.

Of course, the Causal View need not rule out an analysis of causation in terms of counterfactual dependence. When we look for the causes of an effect one natural way to do this is by considering what the effect depended on. This dependency, furthermore, is usually thought of as counterfactual: an act causes at least those effects which would not have occurred had the act not been performed. It is therefore entirely possible to endorse the Causal View and a counterfactual analysis of causation, in which case the difference between the Causal View and the Difference-Making View would be small.

Because the Causal View is compatible with the Difference-Making View it might be wondered whether there is anything substantial at stake here or if the dispute is a mere terminological quibble. In order to bring out the distinctive claim which the Causal View is committed to, let us first note that it is uncontroversial that causation and responsibility are both dependence-relations. To say that *c* causes *e* is to say that *e* depends, in some way, on *c*. Likewise, to say that an act is responsible for an effect is to say that the effect depends, in some way, on the performance of the act. What is distinctive of the Causal View is that according to this view it is the same dependence involved in both causation and responsibility. However, before accepting the Causal View's claim that these relations are the same, we should consider non-causal dependence-relations and how they relate to responsibility.

For example, consider constitution. If *x constitutes y*, then *y* depends on *x* in a non-causal way. Suppose now that the performance of an act constitutes a harmful state of affairs. Examples where an act could constitute a harmful state of affairs are various forms of discrimination and oppression. A particular act of discrimination, such as refusing a red-headed customer, could on some views at least be a case of harming the customer.[20] In such cases, the negative "effect" (the fact that one is subjected to discrimination) is not caused by any individual act but it is constituted by it, so according to the Causal View discrimination as such seems to be relatively harmless. However, what is relevant to responsibility in such a case seems to be whether the state of affairs depends on the performance on the act. Whether this dependence is causal or not is quite beside the point. The Causal View therefore seems to be too narrow.[21] What we should say is that responsibility is often, but not necessarily, causal. Causality can also be relevant to responsibility in other ways. As was suggested above for example, "salience" should be characterised partly in terms of causal relations.

[20] Alternatively, it could be suggested that the discrimination is not constituted by the act alone but by a number of similar acts. In such cases, a particular instance of discrimination would partly constitute a (possibly) harmful state of affairs.

[21] This also shows that it is not a plausible option for a defender of the Causal View to claim that causality should replace talk of responsibility, rather than serve as an analysis of responsibility.

Insisting that the state must depend counterfactually on the performance of the act might be accused of begging the question against the Causal View. It would at any rate strengthen the case against the Causal View if there were examples of properties which causality is typically thought to have but which responsibility is typically thought to lack, or vice versa.[22]

Here there are two properties in particular worth noting.

First, singular causal relations are usually thought to bear some relation to general causal relations, or laws. This is a feature of causal relations which distinguishes them from other dependencies, such as mere correlation for example. However, responsibility does not seem to require that the dependency is general or law-like. Rather, what´matters to responsibility is that a particular effect depends on a particular act. Consider again constitution. That a state of affairs is constituted by an act seems to be enough for us to say that the act is responsible for the state of affairs. Whether this dependence is general, and whether the state of affairs would depend on the act under other circumstances, is beside the point.

Second, causality is usually thought to be a transitive relation: if $c$ causes $e$ and $e$ causes $f$, then $c$ causes $f$.[23] Responsibility, however, is not transitive. An example where responsibility appears to be transitive is the sequence of events connecting Black's pulling the trigger and White's death. Black is responsible for pulling the trigger, and the pulling of the trigger is responsible for the gun firing and so on up to White's death. In such cases, responsibility and causality seem to coincide.

However, there are plausible counterexamples to the transitivity of responsibility. One example are cases where one of the intermediate events is another person's act. Suppose that Purple is responsible for Black's firing of the gun. Perhaps he pointed out White's whereabouts to Black, or perhaps he sold the weapon to Black. Then we would, typically at least, hold Black and not Purple responsible for White's death. In this case it seems reasonable to say that Purple caused, or was at least a partial cause, of White's death. But, even though Purple is responsible, at least to an extent, for Black firing his gun and Black is responsible for White's death, we would not say that Purple is responsible for White's death.

That responsibility can fail to be transitive might seem to be a problem for an analysis of responsibility in terms of counterfactual dependence as well. Here there are two points to keep in mind. First, counterfactual dependence

[22] There might of course be other restrictions on responsibility which are similar to popular restrictions on causality. For example, causation is directed forwards in time (there is no backwards causation) and this is also true of responsibility, at least in the sense discussed here. However, it is enough to show that causation and responsibility differ with respect to some properties for the Causal View to fail.

[23] The transitivity of causation is not completely uncontroversial, though it is a common desideratum for analyses of causation. See for example Hall (2000) and Paul (2000).

is not transitive.[24] Second, *simple* counterfactual dependence as a theory of responsibility is clearly implausible. In order to solve cases like Overdetermination and Pre-Emption we need a more refined theory, for example the Difference-Making View-3. But, this more refined theory can fail to be transitive for other reasons which do not have to do with the counterfactual. It does not follow, for example, that if Purple makes a salient difference to the way in which Black performs his act, then he also makes a salient difference to White's death.

Causation is therefore neither necessary nor sufficient for responsibility. It is not necessary because, as the examples of non-causal dependence-relations show, responsibility need not be "law-like" in the way which causal relations are usually thought to be. Causation is not sufficient because causation is transitive while responsibility is not.[25]

## 5.3 Summary

I have argued that the third condition in the basic structure, the "responsibility condition", should be analysed in terms of counterfactual dependence:

> (3*) if $\phi$ had not been performed then $S$ would not have obtained at all, or would have obtained in a different way which is salient in the circumstances.

Analyses of responsibility in terms of counterfactual dependence faces several well known problems, most notably Overdetermination and Pre-Emption. I have argued that analysing responsibility in terms of counterfactual dependence, more precisely (3*), has acceptable consequences in these cases. According to this view it is not the case that Black is responsible for White's death in Pre-Emption because the effects of his act are never realised. Orange, on the other hand, is responsible for White's death. This, I have claimed, is an acceptable result. In Overdetermination we can claim that both Black and Orange are responsible for White's death because they make a difference to the way in which White dies. The difference they make is also salient in the circumstances which explains why they are responsible while other agents, who make non-salient differences to White's death such as whether he dies in the shade or not, are not responsible. I have also argued that though (3*) is similar to the Counterfactual Condition from chapter two, it is nevertheless a distinct view and that the Counterfactual Condition does not follow from taking (3*) to be necessary and sufficient for responsibility.

[24] See Lewis (1973*b*, pp. 32–35).

[25] A different argument against the Causal View has been offered by Sartorio (2004). Sartorio argues, in short, that when several acts are necessary for an outcome we should say that each individual act is responsible for the outcome but that none of the acts cause the outcome.

Finally, I argued that we should reject the otherwise simpler view that to be responsible for a state of affairs simply is to cause it to obtain. I argued that responsibility and causality differ because there are non-causal dependence relations which are cases of responsibility and because causality is transitive while responsibility is not.

This chapter concludes the characterisation of the basic structure of harm. In the next chapter I will consider whether a complete analysis requires the addition of further conditions or if we should claim that the conditions in the basic structure are not only necessary but also jointly sufficient for doing harm.

# 6. Further conditions

In the last two chapters I have been developing an analysis of harm which takes the basic structure as its starting point. The view I have been arguing for so far can be summed up as follows:

> *a* harms *b* only if
> (1) *a* performs an act, $\phi$,
> (2*) *b* is in a state *S* such that *b*'s life is worse, taking *S* into account, than *b*'s life not taking *S* into account.
> (3*) if $\phi$ had not been performed then *S* would not have obtained at all, or would have obtained in a different way which is salient in the circumstances.

So far I have only taken these three conditions to be necessary for doing harm. In this chapter I will argue that we should endorse a *Minimalist View* about harm which holds that (1), (2*) and (3*) are also jointly sufficient. This view is "minimalist" in the sense that it includes the three conditions from the basic structure but no further conditions.

The way I will proceed is as follows. First, according to the Minimalist View, doing harm consist only in an act being performed, the obtaining of a certain type of state of affairs (a "harmful state of affairs") and the obtaining of the proper relation between the act and the state of affairs. However, there are cases which seem to suggest that this view is too wide and that the Minimalist View has to be complemented with some further condition. I will argue that while some examples suggest that we should make the analysis more narrow by adding further conditions, doing so would not be an overall improvement on the view (sections 1 and 2). The examples which seem to indicate a need for further conditions only support the conclusion that there are factors other than harm which may be important to the *moral* assessment of a person's act, but these further conditions should not be included in an analysis of harm.

Second, I will consider the objection that the Minimalist View is too far removed from common sense in order to serve as an analysis of harm. I will argue that while the Minimalist View departs from common sense in certain respects, this can be explained if we consider that common sense usually does not distinguish between different senses of harm (section 3). Because the Minimalist View is intended to be an analysis of harm in a rather specific sense it is not a fault of the analysis that it does not capture other senses of harm.

## 6.1 Intention and foresight

One suggestion is that we should include a condition pertaining to the agent's intentions. Consider, again, the following case from chapter two:

> *Surgery*. Black saves White's life by amputating White's leg. White suffers intense pain but had Black not amputated the leg then White would have died.

In cases like Surgery the Minimalist View seems to depart from common sense. The pain which White suffers is plausibly something which makes White's life go worse, so it satisfies (2*). The pain is also a state of affairs which is counterfactually dependent on the surgery being performed, so (3*) is also satisfied. Black would therefore harm White by performing the amputation. But, according to common sense we would not typically say that Black harms White in a case like surgery.

Another example where the Minimalist View departs from common sense is where a state of affairs depends counterfactually on an act but where we would not say that the act is responsible for the state. For example, every person's death depends counterfactually on his or her birth. But, we would not say that parents are responsible for their children's death by conceiving them even though a child's death would not have occurred had she not been conceived.

Adding an intention-condition promises to lessen these counter-intuitive implications by saying that an act is responsible for a state of affairs only if the agent *intends* the state of affairs. In Surgery it could then be argued that whether Black harms White depends on whether Black intends the pain or if he intends to save White's life. Suppose that Black only intended to cause White pain but that, unknown to Black, he actually saved White's life by performing the amputation. In that case it seems more plausible to say that Black harms White. However, this conclusion is not as plausible if the purpose of the amputation is to save White's life. Including an intent-condition would allow us to explain the intuitive difference between these two versions of the Surgery case. Similarly, a couple's choice to have a child is only responsible for the child's death if the purpose of the decision is the child's death. Normally this is not the case.

Examples like these seem to favour adding an intent-condition to the Minimalist View. However, including an intent-condition would make the analysis too narrow. For example, the current emission of greenhouse gasses contributes to global warming and will most likely result in a lot of harm due to flooding, malnutrition and disease. This, it seems, is a paradigmatic example of harm being done to future generations. However, these states of affairs are not the intended effects of the emissions but are merely unintended side-effects. Including an intent-condition would however imply that the emission of greenhouse gasses does not harm future generations. Perhaps it could be argued that global warming will not in fact harm future generations in a morally

relevant sense but that it is nonetheless wrong for other reasons. The claim that global warming would not harm future generations is perhaps only counter-intuitive if it also follows that there is nothing morally objectionable about global warming.

As I argued in chapter one regarding other person-affecting approaches to the non-identity problem, this reply is not very promising. It might be the case that some obligations to future generations are not "harm-based" but it does not seem plausible that the harm global warming will bring about is not of the morally relevant kind. After all, the effects of global warming on future generations are paradigmatic examples of harm.

We should also be sceptical about the examples meant to support including an intent-condition because these examples fail to show that intentions are relevant to harm. In Surgery, for example, the reason it seems counter-intuitive to say that there is harm done is because it is easy to think that if an act does harm then there is a serious objection to performing it. However, recall that my aim is to analyse harm as it occurs in the Harm Principle. Saying that amputating would do harm only implies that there is a reason against performing the amputation but nothing about the seriousness of this objection, or whether the person performing the amputation should be blamed for what she does, follows from the claim that there is a reason against performing the amputation. As the Harm Principle was formulated, the only thing that follows from that an act would do harm is that there is *a* reason against performing the act. This is, it seems to me, compatible with our intuitions about a case like Surgery. While it might be odd to say that one does harm by amputating, the important claim is rather that it is *permissible* to perform the amputation and that the person cannot be *blamed* for performing the amputation. Both these claims are compatible with an analysis of harm which does not include an intent-condition. What the Surgery case shows is, at most, that intentions are morally relevant in some way but leaves it open exactly how.[1]

It might be granted that harm is done in cases like Surgery but that we should add some condition in order to exclude cases where the effect is too remotely connected with the act in order for the act to be responsible for it. Adding to the Minimalist View that an act harms a person only if the state of affairs which the act is responsible for is *foreseen* by the agent would exclude such cases. This constraint might seem especially attractive since I have argued that responsibility should be analysed in terms of counterfactual dependence. Without some constraint on this relation states of affairs which would obtain in the far future could determine whether some choice made today does harm. This, it could be argued, is unreasonable unless it is at least possible at the present to estimate whether these effects in the far future will obtain or not.

[1] See for example Scanlon (2008). Scanlon suggests that intentions can be relevant either when one tries to figure out the likely consequences of an act or for determining the "meaning" of an act.

Including a foresight-condition would seem to accomplish similar things which the intent-condition was meant to accomplish while avoiding some of the problems of the intent-condition. For example, the foresight condition does not imply, as the intent-condition did, that there is no harm being done by global warming. But, it still exclude cases where an effect is brought about due to very unlikely circumstances or "freak accidents".

However, if the point of including a foresight-condition is to narrow down the scope of what an act is responsible for then it would not be enough. Take the case of my death being counterfactually dependent on my birth. Parents can surely foresee that their children will, sooner or later, die so the foresight condition would be satisfied in this case. It might be objected that it is not possible to foresee a person's *particular* death, that person's death as it will actually occur, at the time of conception. This is sometime true[2] but relying on this reply would undo the advantage which the foresight-condition had with respect to global warming. We cannot foresee the particular harms which global warming will produce, only that global warming will produce harms of a certain kind to someone in the future.

That the foresight-condition would still be too wide for common sense is of course not a conclusive reason not to include it in an analysis of harm. What it shows is that it does not do what its proponents want it to do. However, the foresight-condition would also make the analysis of harm too narrow because it would make doing harm dependent on the beliefs of the agent. What an agent can foresee in a given situation depends on her beliefs about the situation and the effects of her act. But, we tend to think that an act is responsible for effects which could be foreseen given the facts rather than the agents' beliefs about the facts.[3] For example, if Black points a gun at White and pulls the trigger, then we would hold Black responsible for White's death even if Black believes that White would not die from a bullet because God (or some other powerful entity) would never allow it. Black need not even have these odd beliefs. We would typically hold Black responsible for the effects of firing a loaded gun even in a case where Black lacks a belief about the effects of this act. But the fact that Black in this case cannot foresee that White would die is not a good reason for saying that Black is not responsible for White's death, nor that Black does not harm White.

A proponent of the foresight condition might suggest that the condition should obviously be formulated in terms of what could *reasonably* be foreseen in a given situation. She might agree that taking what people actually are able to foresee would make the condition too narrow. However, it seems fair to ask of the proponent of this revised version of the foresight-condition to spell out

[2] It is possible to foresee a person's death in cases where the person is born with a terminal illness, for example.

[3] Recall that the use of responsibility which is relevant here is not the one which is used when we "hold a person" responsible. In my use of the term, acts (or more generally, events) are responsible for effects in the sense that the effect can be attributed to the act.

what "reasonable" means in this context. This will, probably, involve some claims about what the agent should be able to foresee. On this version of the foresight-condition, whether a state of affairs can be foreseen is not limited to the epistemic situation of a particular agent. White's death can reasonably be foreseen if Black fires a loaded gun at her, even though Black does not have immediate access to this fact, or other facts which would allow him to realise this.

But, modifying the foresight condition in this way makes it unclear whether the reference to what the *agent* can reasonably foresee matters at all. What the foresight-condition amounts to on this revised version is that the effects must be "foreseeable", though not necessarily by the agent. The epistemic situation of the agent seems to drop out almost entirely. Of course, this is not to say that whether White's death is foreseeable might depend on features of the situation. However, whether Black *should* be able to foresee White's death or not can not depend on what Black's actual beliefs about the situation are.

Perhaps this is just what the proponent of the revised foresight-condition was after when formulating the condition. Nevertheless, the condition in its revised form would still be too narrow. Consider the following example from Jackson (1991, pp. 462–3):

> *Dangerous Drugs*. Green, a physician, has to decide which drug to administer to her patient. She has three drugs to choose from: *A*, *B* and *C*. After consulting all the available evidence she knows for certain that *A* will result in a partial cure. She also knows that one of *B* and *C* will result in a complete cure while the other will result in the patient's death, but she does not know which is which.

The general consensus regarding Dangerous Drugs is that objectively Green ought to administer the drug which will result in a complete cure (*B* or *C*) and subjectively she ought to administer *A*. It is then often claimed that the case shows that what Green ought to do, full stop, is the same as what she ought subjectively to do (administer drug *A*). It might therefore seem that this case should only strengthen the case in favour of the foresight-condition since the foresight-condition is also subjective in character. However, on closer examination we can see that the opposite is in fact the case. Consider what we would say about Dangerous Drugs if we ask what Green ought to do if she wants to minimise harm. If we add the foresight-condition to the Minimalist View we can then say the following: Green's choice to administer drug *A* would be responsible for the partial cure because it can reasonably be foreseen what the effects of administering this drug will be. The choice to administer drug *B* (or *C*) would *not* be responsible for the effects of this choice because it is not possible to foresee whether administering drug *B* (or *C*) would result in a complete cure or the patient's death. If she were to administer drug *B*, say, and this would result in the patient's death, the foresight-condition would rule out that Green has harmed her patient, but this seems clearly wrong. There should be no doubt that Green would harm her patient if she administers the drug which

leads to the patient's death, even though she cannot reasonably foresee which drug has this effect.

A more modest version of both the intent- and foresight-conditions would be to say that, other things being equal, it is worse to do harm if one intends (foresees) the harm than to merely do harm without intending (foreseeing) it. However, note that this more modest version does not support the intuitions which motivated the intent- and foresight-condition in the first place. The more modest version implies that Black harms White in Surgery for example and that parents routinely harm their children because they are responsible for their children's deaths.

Still, it might be argued that the modest version of the intent- or foresight-condition captures the plausible claim that there is a morally relevant difference between a case where a bad state of affairs is intended (foreseen) and one where it is not. But, this intuition only supports that intentions (foresight) matters to right, wrong and permissibility, not the claim that the act where the state of affairs is unintended is less harmful.

One example of a view which places a certain weight on intentions and foresight is the so-called "doctrine of double effect". According to this doctrine there is a morally relevant difference between intending harm as a means to an end and merely foreseeing that doing harm is necessary to achieve an end. In the former case there is a reason against doing harm while in the latter it is permissible to do harm. Note however that the doctrine of double effect is ultimately a claim about permissibility and the moral relevance of harm, not about the relevance of intentions or foresight to an analysis of doing harm.[4]

Adding either the intent-condition or the foresight-condition in their modest forms to the Minimalist View therefore seems like a rather desperate attempt to make any difference in the moral status of an act depend on a difference in the degree of harm done, something which we cannot assume to be true.

## 6.2 Consent

Another condition which has been suggested is that an act harms a person only if the person has not given her consent to the act being performed.[5] Consider the Surgery case once again. It might seem plausible to say that performing the amputation would not be to do harm because the patient would consent to the amputation being performed. Cases of self-inflicted injuries or misfor-

[4] There are further differences between the doctrine of double effect as it is usually understood and the the two conditions which I have been arguing against in this chapter. For example, the doctrine is usually only thought to apply if the intended end is good. The intent-condition places no such restrictions on the value of the end. According to the intent-condition, if an effect is not intended then an act which is responsible for the effect does not no harm. See also McIntyre (2001) and Scanlon (2008).

[5] See for example Feinberg (1987, pp. 35, 215).

tunes might also be a reason to include this condition. Some people have the intuition that there is a difference between a smoker who gets lung cancer because of her habit and a non-smoker who gets lung cancer as a consequence of working in a an environment where smoking is allowed. This intuition could be explained by the relevance of consent to harming. The smoker does not harm herself because she exposes herself to smoke willingly, i.e., she has given her consent. In the latter case it could be claimed that the person is harmed because being exposed to smoke, and the related risk of lung cancer, is not something she has given her consent to. Finally, certain cases of bad luck also support including a consent condition. If an act is expected to have a beneficial effect but misfires, and a person suffers a harmful effect as a result, then it might be claimed that the act does no harm. In such cases where the harmful effect is simply due to bad luck then a consent condition would rule out that there has been harm done because of the victim's consent to the performance of the act.

There are several ways to formulate such a consent-condition. Here it is common to distinguish between *actual* and *hypothetical* consent. In terms of actual consent, the condition says that an act does not harm a person if the person has, as a matter of fact, given her consent to the act being performed. The hypothetical version, on the other hand, says that an act does not harm a person if the person *would*, under some suitable circumstances, give her consent to the act being performed. Of these two the hypothetical version seems far more plausible. It is for example unclear why actual consent would matter because people may be badly misinformed about what the consequences of an act would be. Also, the actual version makes a distinction where there does not seem to be a difference. The actual version implies, for example, that there is a difference between performing life saving surgery on an unconscious patient and a conscious one but this seems to be quite irrelevant. The hypothetical view is not vulnerable to such counterexamples and is therefore more plausible than the actual view.[6]

A condition in terms of hypothetical consent would however make the analysis too narrow. Consider what the hypothetical consent-condition would imply with respect to the non-identity problem for example. In the non-identity problem it is assumed that future people would have lives that are well worth living. Given this, would they consent to being brought into existence, even if this meant being born with a "bad start in life"? Because the alternative is

[6] What about actual, informed consent? That is, should we say that it is actual consent that matters, but only when it is informed? This version of the consent-condition would be slightly more plausible than the simple actual version, but not much. The main problem for this version of the condition is that people can consent, or withhold their consent, for irrelevant reasons. For example, a worker in an environment where smoking is allowed might be well informed and might wish that smoking was not allowed but might still give her consent because she wants to remain employed. In such cases, consent is given but for reasons which are clearly irrelevant to whether the person has been harmed.

non-existence these people would have strong reasons based on self-interest for giving their consent. It might be objected that they would also have reasons based on moral principles for not giving their consent. However, if their objection is based on a moral principle which does not allow them to be brought into existence then their consent becomes redundant. In that case it is the principle, not their consent, which does the work since the moral principle would apply to us as well.

It might be suggested that the absence of consent is not necessary for harm but that it only modifies the degree of harm: a person who consents (or would consent) to an act suffers a lesser harm than a person who does not, other things being equal. It should be noted right away that this new version can not be used to capture the intuitions mentioned above, but some still find it plausible that there is a connection between harm and consent.[7] For example, this view could perhaps still account for the intuition that, in general, self-imposed harms carry less moral weight than harm done to other people.

In the previous section I argued against including similar versions of the intent- and foresight-condition and the same point can be made against this version of the consent condition. Because this new version of the consent-condition does not capture the intuitions above, adding the new version to the Minimalist View seems to be an unwarranted attempt to include in an analysis of harm every consideration which seems to be relevant to the permissibility of an act.

We should also be sceptical about the intuitions which seem to support including a consent condition in the first place. It is not evident that they support the claim that consent, whether hypothetical or actual, matters in itself or whether consent is merely indicative of there being further relevant factors. For example, people tend to act with an eye to their own well-being and to avoid harm. In so far as a person perceives a situation correctly we can therefore expect her to consent to acts which will not harm her, at least all things considered. This alternative explanation of the intuitions which appear to support a consent-condition also strikes against the hypothetical version of the consent-condition. The reason we should not take ill-informed consent into account has nothing to do with the consent as such. Rather, the reason it is in practice a good idea to take informed consent into account is because informed consent indicates that there are further factors which are relevant to whether an act is permissible or not without being such a factor itself.

## 6.3 Without further conditions

So far I have argued that we should not add any conditions pertaining to intentions, foresight or consent to the Minimalist View. That is, we should accept the Minimalist View of harm:

[7] See for example Harman (1981, p. 293) and Shiffrin (1999, p. 130)).

*a* harms *b* if and only if
(1) *a* performs an act, $\phi$,
(2*): *b* is in a state *S* such that *b*'s life is worse, taking *S* into account, than *b*'s life not taking *S* into account.
(3*) if $\phi$ had not been performed then *S* would not have obtained at all, or would have obtained in a different way which is salient in the circumstances.

In chapter one I introduced three desiderata which an analysis of harm should satisfy. It should solve the non-identity problem, be compatible with the no-difference view and be intuitively acceptable. The Minimalist View of harm fares well with respect to the first two of these desiderata. First, in the case of the young girl it follows that she would harm her child if she does not postpone her pregnancy, provided that "a bad start in life" satisfies (2*). This proviso is clearly fulfilled because we should understand "a bad start in life" to mean at least that the life is worse for the person, taking the bad start into account, than not taking the bad start into account.

Second, the analysis is consistent with the no-difference view. The two medical programmes would be equally worthwhile because they would both prevent an equal amount of harm, again provided that the handicaps which the programmes would prevent satisfy (2*). Note however that because the handicaps in the example are supposed to be qualitatively identical there is no reason to think that one condition would satisfy (2*) while the other would not. That is, either both prevent harm or neither does. This is perfectly consistent with the no-difference view and merely reveals that the exact implications of the analysis will depend on what theory of well-being is "plugged in".

Regarding the third desideratum, that an analysis of harm should be intuitively acceptable, it must be granted that the Minimalist View amounts to a wider notion of harm than is commonly accepted. The reasons for including further conditions, and thereby making the analysis more narrow, are as I have argued above not very compelling.

An objection at this point is that my view makes harm so far removed from the ordinary usage of the term that it does not qualify as an analysis of harm in any sense that matters. The conclusion that we should not add any further conditions to (1)-(3*) does not in itself show that we should endorse the Minimalist View, that is, that these conditions are necessary and jointly sufficient. An alternative conclusion would be that harm is a heterogeneous concept which it is not possible to give necessary and sufficient conditions for. The upshot of this objection is that we should abandon harm as a morally relevant notion.

In the remainder of this chapter I will consider a version of this objection which has been advanced by Bradley (2012). Bradley argues that harm is too diverse a concept to be of any use in serious theorising and that it should be replaced by more well-behaved concepts, such as intrinsic and extrinsic value.

In reply, I will argue that while my analysis departs from common sense, it still serves as an acceptable analysis of harm in a specific sense, namely as the

term occurs in the Harm Principle from chapter one. Because the Minimalist View is not an analysis of harm in a wider sense it is not a failure of the analysis that it does not capture intuitions about harm which only concern this wider sense.

### 6.3.1 Bradley's objection

Bradley (2012) argues that harm is a "Frankensteinian" concept which is too diverse to be of any use in serious theorising. He argues for this conclusion by listing a number of desiderata which an analysis of harm should satisfy. But, because every available analysis fails to satisfy all of these desiderata, Bradley concludes that

> harm is a Frankensteinian jumble. Thus it is unsuitable for use in serious moral theorizing. It should be replaced by other more well-behaved concepts, such as the axiological concepts of intrinsic and extrinsic badness. (Bradley 2012, p. 391).

In order for the Minimalist View to be acceptable, we therefore have to consider how it fares with respect to Bradley's desiderata and whether the desiderata are reasonable in the first place. As we will see below, my analysis fares well with respect to several of Bradley's desiderata, but it stumbles on some. However, I will argue that this is not sufficient to reject the analysis. We need to keep in mind that the aim in this thesis is to analyse harm in a rather specific sense. The purpose of the analysis, which can be identified with its normative role in this case, has to be specified at the outset and an analysis is successful in so far as the purpose can be achieved without departing too much from ordinary usage.

Bradley suggests that an analysis of harm should satisfy a total of seven desiderata: (i) extensional adequacy, (ii) axiological neutrality, (iii) ontological neutrality, (iv) amorality, (v) unity, (vi) prudential importance and (vii) normative importance. Of these seven, my analysis satisfies (ii)-(v) and I will therefore only briefly indicate why that is the case.[8] First, my analysis is axiologically neutral because I have not made any substantive assumptions about well-being (see chapter four). Second, while the Minimalist View I favour is not ontologically neutral as it stands, it can easily be modified. I have formulated the analysis in terms of acts and states of affairs but nothing hinges on this choice. For example, (2*) could be formulated in terms of events, or even a mix of ontological categories. Third, the analysis is "amoral" because it does not entail that harming is morally wrong or objectionable, that doing harm requires malicious intent nor that harm-claims entail a deontic judgement. Fourth, the analysis is "unified" because it is not merely a list of typical examples of harm.

[8] For a complete description of these desiderata, see Bradley (2012, pp. 394–6).

This leaves three desiderata: extensional adequacy, prudential importance and normative importance.

Regarding extensional adequacy, Bradley claims that "the analysis must fit the data. [...] If no analysis gets all the data right, we should favor the one that does better by the data, all else equal" (Bradley 2012, p. 394). This desideratum seems obvious, and also seems to be the main point where my analysis fails. For example, I have argued (see chapter four) that we should interpret the second condition in such a way that absences of benefits satisfy the condition.

However, there are several things to note about extensional adequacy as a desideratum. First, as Bradley seems to be aware, it would be to expect to much of an analysis that it should fit the data perfectly. Rather, we should strive for the *best* fit. That my analysis departs from the data in certain cases does therefore not show by itself that it should be rejected. Second, there is considerable disagreement about what the data are when it comes to harm. As we have seen in previous chapters, there is disagreement about whether failing to benefit a person for example, is to do harm.

Another source of disagreement is the distinction between partial and total harm. Cases where an act makes a person better off all things considered but where the act involves some misfortune are also cases where people tend to disagree about whether the act does harm or not. If there is no agreed set of data then we should not insist that an analysis should fit the data.

Furthermore, the examples above which seem to suggest that the Minimalist View departs significantly from common sense are not decisive. As I mentioned in chapter two, it is important to distinguish between *total* and *partial* harm. The former concerns the total state of affairs, how things are for a person all things considered. The latter merely concerns how a certain effect, or a certain act, affects a person. Judgements about partial harms are not judgements about how things are for a person all things considered. The claim that life-saving surgery, for example, is a form of harm is clearly only plausible as a claim about partial harms. In saying that the surgery harms the patient one is certainly not saying that the surgery was morally forbidden, one is merely saying that there was something about the surgery which was bad for the patient and that this should be taken into account when doing a moral evaluation. While the account is revisionary, it does not contradict common sense because common sense usually overlooks a distinction which is relevant on this account.

The fact that there is little agreement about the data might seem to undermine the project of giving an analysis of harm in the first place. However, we can still analyse harm in more specific senses. The aim of this thesis has been to give an analysis of harm which makes the Harm Principle plausible, especially with respect to same-number cases and the non-identity problem. Whether it succeeds or fails should therefore be judged by how well it captures harm in this sense, not some other.

Judging whether an analysis satisfies extensional adequacy independently of its normative role is for the reasons just given problematic. We should therefore take a closer look at Bradley's last two desiderata – prudential and normative importance – before deciding whether the Minimalist View passes as an extensionally adequate analysis.

Regarding prudential importance Bradley claims that "harm is something worth caring about in prudential deliberation. Harm is the sort of thing we should try to avoid; if we have an analysis of harm such that one might reasonably be indifferent concerning whether an event of the sort in the *definiens* takes place or not, we should reject the analysis" (Bradley 2012, p. 395). My analysis has an obvious connection with reasonable prudential concerns because it is formulated partly in terms of well-being. All harms, on my view, make a person's life go worse and reasonable prudential concern clearly involves concern for one's own well-being.

On the other hand, it might be argued that an act could "harm" a person according to my analysis but the victim could be indifferent, or even glad, that the act was performed. For example, making a person experience some lesser pain in order for that person to avoid greater future pain, or to secure great future benefits, would on my analysis be to do harm. But, the "victim" in such a case can be quite satisfied with experiencing the lesser pain.

In defence of my view it can be pointed out that it still makes sense to care about the lesser pain. It still seems reasonable to regret that one has to undergo the lesser pain, and one can reasonably resent that one has to undergo it in order to avoid the greater pain or secure the future benefit. If given a choice between the lesser pain as a means to a greater benefit and the greater benefit without the lesser pain, then it is certainly reasonable to prefer the latter. Harm, on my analysis, is therefore something which it makes sense to care about in prudential deliberation, even though it may not matter quite as much as Bradley seems to require.

Finally, regarding normative importance, Bradley claims that

> the analysis should entail that harm is the sort of thing that it makes sense for there to be deontological restrictions about. If an analysis of harm, when plugged into Mill's harm principle or one of Frances Kamm's deontological principles, makes the principle absurd on its face, then it is not what we are looking for. (Bradley 2012, p. 396).

My analysis would indeed make some deontological principles absurd, or at least very implausible. For example, a principle which stated that it is *forbidden* to do harm would be quite absurd given my analysis because it would make too many acts forbidden. Another example, though it does not qualify as a deontological principle perhaps, is the Hippocratic oath. Taking an oath to do no harm would, given my analysis, be quite absurd because it would be nearly impossible to keep it.

However, this desideratum seems to be too strong. Recall that the analysis of harm which I have been arguing for is an analysis of *partial* harm, not *total* harm. The deontological principles which turn out to be very implausible given my analysis are, I propose, those principles which appeal to harm in a total sense. Indeed, it seems that *any* analysis of harm in the partial sense will make a deontological principle which forbids doing harm absurd. The Hippocratic oath, for example, would be quite implausible if "harm" is understood in the partial sense. But this cannot be a good reason for disqualifying analyses of partial harm across the board. A weaker deontological constraint, such as the Harm Principle which was introduced in chapter one, would not be absurd on its face given my analysis.

A further reason for thinking that Bradley's desideratum is inadequate as it stands is that normative principles which refer to harm are not "data" which the analysis should account for. As I argued in chapter two, an analysis of harm should be judged by its descriptive adequacy and its normative adequacy. So far I am in agreement with Bradley. Descriptive adequacy is, obviously, important because if an analysis of harm is not able to account for many harm-claims in ordinary talk then we seem to have made harm into a technical term with little or no connection to ordinary use. However, as I argued above, whether an analysis is descriptively adequate cannot be judged independently of its normative role. Normative and descriptive adequacy go hand in hand and we cannot determine whether an analysis succeeds descriptively unless the normative role has been specified. Normative, or theoretical as the case may be, adequacy is achieved when the analysis serves its purpose. That is, when it fits the role it is intended to have.

For these reasons, Bradley's desideratum concerning normative importance is too demanding. A more plausible way to formulate the desideratum is that an analysis of harm should be such that it would not be absurd to claim that harm is the sort of thing which matters morally in some way. The normative role of my analysis is fairly specific. I set out to analyse harm *as it occurs in the Harm Principle*. This is to analyse harm in a specific sense. For example, it is an analysis of harm in the partial sense as opposed to the total sense. It is therefore limited in its application and should not be taken as an analysis of harm as it occurs in other contexts, such as the Hippocratic oath. To draw the conclusion that this limitation disqualifies it as an analysis of one sense of harm would be a mistake.

#### 6.3.1.1 Does harm matter?

In short, my reply to Bradley's objection is that the Minimalist View for should be understood as an analysis of harm in a certain sense, namely as it occurs in the Harm Principle. There are certainly other senses of harm, such as the one used in the Hippocratic oath, but harm in this sense has not been the target of my analysis. Bradley's desiderata are in this respect too demanding.

This reply raises a further worry however. By narrowing the scope of the analysis it might be wondered whether we have not effectively devalued harm and if there is any reason to appeal to this narrow sense of harm in serious theorising. Put in another way, if the Minimalist View is an analysis of harm in a specific sense, why should we care about this specific sense? This is not so much a worry about whether the analysis is acceptable or not but whether there is any particular theoretical gain to be had by appealing to harm in this narrow sense. Perhaps we are better off focussing on well-being, or as Bradley suggests, intrinsic and extrinsic value.

Well-being and different kinds of value are obviously very important to my analysis and the answer to many substantive questions hinges on what the correct theory of well-being is, for example whether absences of benefits are harms. But, this fact does not make the analysis redundant or uninteresting. The analysis contributes to our pre-theoretical understanding of harm by distinguishing different components which harm and harming depend on, such as well-being and responsibility, without taking a stand on how these components should be understood more precisely.[9] Understanding well-being is obviously very important but well-being cannot replace harm.

Of course, even supposing that the analysis contributes to our pre-theoretical understanding of harm does not show that harm has any significant role to play in for example ethics. We do not "have to" appeal to harm merely because we can. If there is no distinct advantage of appealing to harm – if there is no distinct role for it to play – in addition to other common ethical concepts then we have some reason because of parsimony to get rid of harm. In the following chapters I will argue that there is such an advantage to be had.

## 6.4 Summary

In this chapter I have argued that we should not add any further conditions pertaining to intentions, foresight or consent to the basic structure. Certain examples seem to suggest that the analysis would be more plausible by adding any of these conditions but I argued that this is merely apparent. On closer examination, these further conditions would make the analysis too narrow. I also argued that the examples which seem to support including any of these conditions can be explained in other ways. Our tendency to think that intentions, foresight or consent matters to an analysis of harm can be explained by a more general tendency to think that these three notions are relevant to the moral evaluation of an act. However, we should not conclude that they are therefore relevant to an analysis of harm.

[9] It is a mark of a successful analysis, it seems to me, that the need to appeal to the analysandum seems to disappear. That the need for using harm seems to disappear should be taken as an indication that the analysis has been fairly successful.

By not adding any further conditions to the basic structure we are left with the following Minimalist View of harm:

> *a* harms *b* if and only if
> (1) *a* performs an act, $\phi$,
> (2*) *b* is in a state *S* such that *b*'s life is worse, taking *S* into account, than *b*'s life not taking *S* into account.
> (3*) if $\phi$ had not been performed then *S* would not have obtained at all, or would have obtained in a different way which is salient in the circumstances.

I argued that the Minimalist View fare well with respect to two desiderata from chapter one: it solves the non-identity problem and it is compatible with the no-difference view. Regarding the third desideratum, that the analysis should be intuitively acceptable, I argued that while the Minimalist View departs from common sense in some cases, this departure can be explained by appealing to the distinction between partial and total harm.

Finally, I considered a number of desiderata which a plausible analysis of harm should satisfy according to Bradley. The Minimalist View fares well with respect to the majority of these desiderata. Regarding the two desiderata which seemed to be especially troubling for the Minimalist View, extensional adequacy and normative importance, I argued that whether an analysis satisfies extensional adequacy cannot be judged independently its intended normative role. Because the Minimalist View is an analysis of harm as it occurs in the Harm Principle it is therefore not a fault of the analysis that it does not capture other uses of "harm".

My reply to Bradley's objection left us with a problem however. If the Minimalist View is an analysis of harm in a specific sense, should we appeal to harm in *this* sense when doing ethics? The task of the next two chapters I will argue that there is.

# 7. The asymmetry

In this thesis I have been arguing for a Minimalist View of harm. To harm someone, according to this view, is to perform an act which is responsible for a state of affairs (or an event) which makes someone's life go worse. The Minimalist View of harm is an analysis of harm in a certain sense, namely as it occurs in the Harm Principle:

> *The Harm Principle*: if an act would harm someone then this is a reason against performing that act.

In conjunction with this normative principle, the Minimalist View allows us to say that the young girl's choice not to wait would harm her child and that this is a reason for her to wait. Likewise, the Minimalist View is compatible with the no-difference view because it treats same-number cases in the same way as same-people cases.

What I have argued for so far is then merely that the non-identity problem does not rule out the Harm Principle as a relevant normative principle in population ethics. It could be argued that while it is true that one can do harm even in same-number cases, this does not show that it is the Harm Principle which explains what one ought to do in such cases. An alternative solution to the problem is to appeal to *Q* and say that people ought, other things being equal, do what would be best. As far as the non-identity problem goes at least, we do not need to appeal to harm in order to justify our intuition that it is morally objectionable to create the person with lesser well-being. The worry, simply put, is whether we *need* the Harm Principle.

Parfit argues along these lines for what he claims explains what we ought to do in all same-number cases, *Q*:

> If in either of two possible outcomes the same number of people would ever live, it would be worse if those who live are worse off, or have a lower quality of life, than those who would have lived. (Parfit 1984, p. 360)

Parfit claims that *Q* is superior to views that involve a "person-affecting" component to the effect that our choice must be worse *for* someone in order to be worse *simpliciter*. The reason why appealing to *Q* is superior to person-affecting views, such as the Harm Principle, is that while *Q* has the same implications as person-affecting views in same-people cases, person-affecting views do not have plausible results in same-number cases. We should there-

fore conclude, according to this argument, that we should appeal to *Q* rather than person-affecting views. The same line could be pursued against the relevance of the analysis of harm presented in this thesis. On my view, the Harm Principle cannot account for the intuition that it is morally objectionable to create a person with less well-being if there is no harm involved. *Q* is not restricted in this way and gives the same result in non-identity cases regardless of whether there is harm involved or not.

A possible reply to this argument is that one could appeal to both the Harm Principle and *Q*. One could say that when harms are not involved then the reason we ought to have the child who would be better off is because of *Q*. When harms are involved we have an additional reason based on the Harm Principle. The problem with this reply is that once one accepts the appeal to *Q* as a solution to the non-identity problem then there seems to be no need for any further moral principle in order to explain what one ought to do in same-number cases. The Harm Principle becomes an explanatory fifth wheel. We have not been presented with any reason to accept the Harm Principle as an addition to *Q*, while we do have a reason to appeal to *Q* because it gives the intuitively correct result in same-number cases in general.

In this chapter I will argue that views which are congenial with *Q* has counter-intuitive consequences when we consider creating new people. A common intuition, often called "the asymmetry", is that there is a moral difference between making people happy and creating happy people.[1] In section one I will clarify the content of this intuition and why it is significant. In section two and three I will argue that attempts to explain the asymmetry which are congenial with *Q* fail. Only appealing to *Q* therefore comes with a significant cost because it would require revisions with respect to some of our deeply held views about our obligations to future people.

## 7.1 Formulating the intuition

A common intuition is that our obligations to future people are asymmetrical in the sense that, on the one hand, we ought not to bring a person into existence if she would suffer intense pain throughout her whole life. On the other hand it is not the case that we ought to bring a person into existence if she would live a happy life. When it comes to having children, it seems to be morally objectionable to have them if they would suffer but morally neutral to refrain from having them if they would be happy.

Even though this asymmetry is rather intuitive many find it in need of a justification. Normally, if an act would lead to someone being (very) happy then this counts in favour of performing that act. Why is this not so when it comes to creating future people? If there are no serious considerations against

[1] Narveson famously expressed this intuition with the slogan "[w]e are in favor of making people happy, but neutral about making happy people" (Narveson 1973, p. 80).

creating a person who would live a happy life, should we not say that it is morally required to create that person? That there is a morally relevant difference between making existing people happy and making future people happy seems to be *ad hoc*. My aim in this chapter is to argue that typical approaches to the asymmetry fail. In the next chapter I will argue that we can do a much better job of justifying this intuition by giving a special role to harm in moral theorising.

First, however, there are a couple of clarifications regarding the intuition itself which should be made. First, the asymmetry is an intuition about future people whose existence depends on our choices. A person whose existence depends on the performance of an act will for the remainder of this chapter be referred to as a "contingent person". A person who will exist whether or not an act is performed will be referred to as a "necessary person".[2] With this terminology we can then say that the asymmetry is a view about benefits to *contingent* future people. It is silent about benefits to *necessary* future people. The asymmetry is therefore compatible with the claim that in a situation where we can benefit some future person, and this future person will exist regardless of what we do, then we ought to do what will benefit this person most.

It is also necessary to clarify the intuition itself and how the asymmetry should be formulated. One way to formulate it is in terms of *duties*.[3] The asymmetry in terms of duties holds that there is no duty to create a person whose life would be worth living, but there is a duty *not* to create a person whose life would be not worth living. It is of course *permissible* to create a person whose life would be worth living. What the asymmetry claims is merely that it is not *required* by morality to create a person merely because her life would be worth living.

This version seems unnecessarily strict however. We should not deny that there can, in extreme circumstances, be a duty to create a person with a life worth living. Suppose, for example, that an evil demon threatens to make every existing and future person's life not worth living unless we see to it that a person with a life worth living is created. Here it would clearly be a mistake to say that we do not have a duty to create a person.

In order for the asymmetry in terms of duties to be plausible its defenders should therefore make the weaker claim that there typically is no duty to create happy people. An example of this way of formulating the asymmetry is suggested by Narveson (1967, p. 66). Narveson claims that there can be a duty to create "happy people" but that in such cases what makes it a duty will not be because of the act's "direct effects" but rather its "indirect effects". One way to interpret Narveson here is as claiming that the duty to create a person

[2] The terms "contingent" and "necessary" people are sometimes used in a wider sense where a person is "contingent" relative to a set of possible worlds if and only if she exists in some but not all of these worlds. A person is "necessary" relative to a set of possible worlds if and only if she exists in all of them. See for example Österberg (1996, p. 100).

[3] See for example Narveson (1967, pp. 65–6) and Elstein (2005, p. 49).

who would have a life worth living can not be grounded in the fact that the person would have a life worth living. Whether there is a duty to create such a person depends only on how the act affects other people, not on how the act affects the person being created.

While formulating the asymmetry in terms of duties may come with some intuitive advantage, it is not precise enough. As Narveson's manoeuvre with "direct" and "indirect" effects suggests, we need finer distinctions than talk of simple duties will allow to capture the asymmetry. The asymmetry is not the intuition that it can never be a duty, under any circumstances, to create a person with a life worth living. What the intuition is about is rather that it is neutral to create people who would have lives worth living taking only their well-being into account.

An alternative to the duties-formulation is to formulate the asymmetry in terms of reasons. The idea is that if a person would be happy if she were to be created then this is not a reason for, it does not count in favour of, creating her. On the other hand, that a person would be unhappy if she were to be created is a reason against creating her. Broome has captured this idea neatly:

> If a person could be created, and would lead a good life if she was created, the fact that her life would be good is not a reason for creating her [...] If a person's life would be bad, were she to be created, that is a reason against creating her; a person's existence is ethically neutral only if her life would be good. (Broome 1999, p. 228).

Formulating the intuition in terms of reasons captures the finer distinctions which a formulations in terms of duties did not. Broome's formulation is, for example, compatible with there being a reason to create a person who would live "a good life" if creating this person would bring great benefits to other people.

The way I will understand the asymmetry is in many ways similar to Broome's formulation. More precisely, I will understand the asymmetry as consisting of two claims about how what is good for future people relates to what we have reason to do:

> (I) If a contingent future person would have a life not worth living then this is a reason against creating that person.
> (II) If a contingent future person would have a life worth living then this is not a reason in favour of creating that person.[4]

By a reason I mean a *pro-tanto* reason; that is, a consideration which counts in favour of, or against, a certain act but which can be outweighed, or defeated, by other reasons. By "a life worth living" I mean a life where the good things in that life outweigh the bad things. By "a life not worth living" I mean a life

[4] For similar formulations of the asymmetry, see McMahan (2009) and Roberts (2011). For an early formulation of the asymmetry in terms of reasons, see McMahan (1981, p. 100).

where the bad things outweigh the good things.[5] We can think of these good and bad things as well-being components: states of affairs, events or things which affect how well a life is going for the person whose life it is. Finally, I will assume that a person whose life is worth living has *positive* well-being. Likewise, a person whose life is not worth living has *negative* well-being.

Formulating the asymmetry in terms of reasons does not mean that whether we accept the asymmetry is of no consequence for our duties to future people. By formulating the asymmetry in terms of reasons we are, however, allowing for some differences regarding how reasons relate to duties. For example, some people hold that if a person has a duty to perform an act, then the person has most reason to perform that act.[6] If one also holds *this* view about the relation between reasons and duties, then we have a possible *explanation* of why it is not the case that we have a duty to create people with lives worth living, other things being equal. We need not, at this juncture, settle for a view on the relation between reasons and oughts. The claim that the well-being a contingent future person would have is not a reason to create her, i.e., does not count in favour of creating her, is sufficiently intuitive to be worth exploring though I will return to the relation between reasons and duties in the next chapter.

### 7.1.1 Strong and weak asymmetry

A distinction is sometimes made between *weak* and *strong* asymmetry.[7] According to strong asymmetry, the fact that a contingent future person would have a life worth living is *no* reason to create that person. The asymmetry as I defined it above is therefore a version of strong asymmetry. According to weak asymmetry, on the other hand, the fact that a contingent future person would have a life worth living gives us *some* reason to create that person. According to weak asymmetry, reasons from benefits to contingent future people provide reasons but they should be discounted; they carry less moral weight than other benefits. These reasons do not, it is then often claimed, amount to a duty to procreate.[8]

The weak version of the asymmetry, that we should discount the benefits to contingent future people, is interesting in its own right but it is not a plausible

[5] For completeness sake we should perhaps also say that a neutral life is a life where the good and bad things are evenly matched, or if there is a complete lack of both. I will not consider neutral lives in what follows however.

[6] See for example Stroud (1998). According to Stroud, this claim captures the "the overridingness" of morality.

[7] See for example McMahan (2009) and Arrhenius (2000, p. 137).

[8] This can be argued for in a number of ways. For instance, some think that it can be accounted for by giving reasons to benefit a lesser weight than reasons against doing harm in general. See for example Harman (2004). Another proponent of this view is W. D. Ross who says that "non-maleficence is apprehended as a duty distinct from that of beneficence, and as a duty of more stringent character" (Ross 2002, p. 21).

replacement for the strong version of the asymmetry. For example, because the weak version assigns some weight to benefits to contingent future people it allows for certain trade-offs which the strong version does not. According to the weak asymmetry, the well-being a future person would have provides us with some reason to create her and therefore it seems to be a consequence of this view that if the benefit is large enough then this reason could outweigh reasons to provide smaller benefits to existing people, other things being equal. In other words, on the weak version it is possible that instead of benefiting existing people we ought to add more people to the world as long as they have lives worth living. This seems counter-intuitive.[9]

The strong asymmetry, on the other hand, holds that there is an important difference between benefiting existing people and creating contingent future people who would have lives worth living. The fact that a contingent future person would have a life worth living is not a reason to create them but the fact that some non-contingent person (whether future or not) would benefit from an act is a reason to perform that act. It does therefore not allow for trade-offs in the same way as the weak asymmetry does. It is therefore worthwhile to see if the strong asymmetry can be captured by views which are congenial with *Q*.

## 7.2 Impersonal axiologies

There are several ways one can approach the asymmetry. One way is to attempt to justify it on axiological grounds and claim that while adding a person with a life worth living does not make the population *better*, adding a person with a life not worth living makes the population *worse*.[10] On an axiological approach one assumes a connection between values and reasons and that the reasons in question are matched by a difference in the intrinsic value of

[9] See McMahan (2009). He lists four consequences of weak asymmetry which he finds "very difficult to believe" and notes that strong asymmetry "may be the only view that captures our strongest intuitions about the morality of procreation" (McMahan 2009, p. 67).

[10] By a "population" I merely mean a collection of individuals and a distribution of well-being among those individuals. On the axiological view, it makes sense to talk about the value of a population, and to say that some are better than others. I will not make any assumptions about *what* the value of a population depends on, though it seems very plausible that the distribution of well-being is important in this respect. Note though that the axiological approach is compatible with the view that the value of a population depends on other properties than just the amount of well-being in that population. For example, it is compatible with the axiological approach to say that an *equal* distribution of well-being makes a population better, other things being equal. It is also important, on this view, to distinguish between the personal value of a life (how good a certain way of life is for a person) and the "contributive value" of a life (the difference a certain life makes to the impersonal value of a population). See Arrhenius (2000, p. 7) and Broome (2004, p. 65). This distinction is analogous to the one made in chapter four where the value of a state of affairs for a person was distinguished from the contribution the state of affairs makes to a particular life.

a population. On this approach, therefore, when one says that there is a reason against creating people with lives not worth living one has to support this claim with a theory of value which entails that adding such a person to a population makes it worse.

An immediate problem for the axiological approach is how to characterise the addition of a person in a positive way. The way the axiological view interprets the intuition is by saying that adding a person with positive well-being does not make the world better, but it might well be asked what difference, if any, it does. Saying that it makes the world worse to add a person with a life worth living seems out of the question, so perhaps one should say that it makes no difference. That is, perhaps we should interpret the asymmetry in terms of equal value: adding a person with a life worth living to a population does not make the population better or worse, its value stays the same.

Broome (2004) discusses (and rejects) an axiological attempt to save the asymmetry along these lines. Broome suggests that there is a range of well-being levels such that adding a person at one of these levels does not make the world better or worse, adding a person below this range makes the world worse and adding a person above this range makes the world better. According to this *intuition of neutrality*, as Broome calls it, "adding a person to the world is very often ethically neutral" (Broome 2004, p. 143).

In order for Broome's view to support strong asymmetry we also need to assume that the neutral range does not have an upper limit. If there were a limit then adding a person above this limit to the world would make it better, but this would of course contradict strong asymmetry. In what follows I will therefore assume that the neutral range does not have an upper limit.

Broome (2004, pp. 146–7) argues that interpreting the intuition of neutrality in terms of equal value is deeply problematic. Suppose that we have three alternatives: (*A*) create no-one, (*B*) create a person who would have a well-being of 10 or (*C*) create the same person but with a well-being of 5. Let us also assume that 5 and 10 are within the neutral range and that our choice will not affect anyone else in the population. Now, if neutrality is interpreted in terms of equal value then it implies that *A* is equally as good as *B* and *C*. However, *B* is clearly better than *C* because *B* is better for someone and at least as good for everyone else.[11] But, if *B* is better than *C* then *A* cannot be equally as good as *both B* and *C*. To say that a person's existence is "neutral" can therefore not be interpreted in terms of equal value.

[11] I here assume what I take to be a rather uncontroversial principle, at least for same-people cases, which Broome refers to as "the principle of personal good" (Broome 2004, p. 58). According to this principle if *A* is better than *B* for someone and at least as good for everyone else, then *A* is better than *B*. This principle rules out one possible way of defending the intuition, namely the view that the value of a population only depends on the amount of suffering, or negative well-being, it contains. However, this view is very counter-intuitive, as the fact that it is incompatible with the principle of personal good illustrates.

As Broome formulates the intuition, that adding a person within the neutral range does not make the world better or worse, does however not entail that neutrality should be understood as having equal value. Incomparability and indeterminacy are two possibilities which we should consider.

That two things are *incomparable* is often defined simply by saying that neither is better than the other and that they are not equally good. Typical examples of purported incomparability are things which exemplify radically different values, like in Sartre's famous example where a student faces a choice between fighting for his country and taking care of his old mother. Neutrality understood as incomparability amounts to the view that, other things being equal, a state of affairs $S$ and the same state of affairs but with one additional person with a life worth living, $S'$, are incomparable.[12]

To say that it is *indeterminate* what value-relation holds between two things is to say that there is no fact of the matter whether one is better than the other or whether they are equally good. This is the alternative Broome opts for and his ground is that it seems plausible to hold that betterness is vague. That is, there is a grey-area of things (in this case, populations) where there simply is no fact of the matter which is better than which.

A problem for both views which Broome discusses, but which I will merely mention, is that they seem to be poor interpretations of neutrality. On both views, Broome argues, neutrality becomes "greedy" in a way that "is capable of swallowing up badness or goodness and neutralizing it" (Broome 2004, p. 170). To see how this argument works we consider the following distributions of well-being:[13]

$$A = (4, 6, -)$$
$$B = (4, 6, 1)$$
$$C = (4, 4, 4)$$

Suppose that the addition of the person in $B$ is neutral, so $B$ is neither better, worse nor equally as good as $A$. Broome also assumes, plausibly, that $B$ is worse than $C$. $C$ has a higher total and average well-being *and* it is more evenly distributed. Now consider $A$ and $C$. $C$ can not be worse than $A$ because, if it were, $B$ would be worse than $A$. If $C$ were worse than $A$ then, because $B$ is worse than $C$ and "worse than" is transitive, it would follow that $B$ is worse

[12] See for example Österberg (1996, p. 100–1) who suggests a view which implies that many cases involving addition of people with lives worth living results in incomparability. Rabinowicz (2009) argues in favour of this approach to the intuition of neutrality. It should be noted however that Rabinowicz does not discuss the strong version of the intuition according to which it is sufficient to add a person with any amount of positive well-being in order for $S$ and $S'$ to be incomparable. Brown (2011) seems to endorse the strong version of this view.

[13] Distributions of well-being are here represented by vectors. A number represents that a particular person exists with that much well-being and a dash represents that a particular person does not exist.

than *A*. But this runs contrary to the assumption that the mere addition of one person in *B* is neutral.

That *C* is not worse than *A* is unintuitive according to Broome. The difference between *C* and *A* is that in *C* one person is added, which is a neutral thing, and one person loses some well-being (the second person in this example) which is a bad thing. Now, Broome thinks that "the net effect of one bad thing and one neutral thing should be bad. But according to our theory, it is not bad; it is neutral" (Broome 2004, p. 170). So *C* should be worse than *A*, according to this argument.[14]

A final objection to both incomparability and indeterminacy is that both seem less plausible as interpretations of strong asymmetry. Recall that the strong version of the asymmetry requires that there is no upper bound to the neutral range. However, both views would still have to maintain that adding a person below a certain level of well-being is not incomparable or indeterminate. This value asymmetry looks very suspicious. What is it about adding people with a life worth living which makes it so different from adding people with lives not worth living?

Consider the case of indeterminacy first. It seems plausible that there is a finite range of well-being levels where our concept of betterness is vague. This fact about our concept of betterness could possibly explain weak asymmetry: there is a range of well-being levels where it is indeterminate whether adding a person at these levels to a population makes the population better. However, the indeterminacy-approach seems much less plausible when it comes to strong asymmetry. In order to capture strong asymmetry one would have to claim that betterness is vague for a certain level of well-being (where the neutral range begins) and for any level above. The only reason to think that the concept would be vague for adding people with lives worth living seems to be that it is the addition of people as such which makes two populations indeterminate with respect to betterness. But, this reason would of course undermine the claim that adding people with lives not worth living makes a population worse; adding such people should, on this view, also be indeterminate.[15]

[14] A weakness in this argument has been pointed out by Rabinowicz (2009). Rabinowicz argues that Broome's greediness objection assumes that if something is of neutral value then it does not count against other values. However, Rabinowicz claims, this is not how we should understand the neutrality intuition. Rather, we should say that "adding people is (axiologically) neutral simply means that it on its own makes the world neither better nor worse. This does not imply that such changes don't "count against other values" and that they can simply be ignored in the total evaluation of outcomes" (Rabinowicz 2009, p. 399).

The point here is that Broome is not entirely consistent in his use of "neutral". If neutrality is to be understood as incomparability or indeterminacy then Broome's claim that "the net effect of one bad thing and one neutral thing should be bad" is false. Rather, if neutrality is understood as incomparability then the net effect of a neutral and a bad thing *should* be incomparability.

[15] The indeterminacy-approach to the asymmetry would also make the vagueness of betterness very different from other examples of vagueness. Typically, if it is vague whether *x* is more *F* than *y* then *x* and *y* are both within a grey area where it is indeterminate whether one is more

A similar argument can be given against incomparability. Interpreting the strong asymmetry in terms of incomparability amounts to the claim that there is a difference with respect to comparability between adding a person with a life worth living to a population and adding a person with a life not worth living to a population. However, one might well wonder why there should be such a difference. For example, two populations of different sizes do not exemplify radically different values, as incomparable things are usually thought to do. Perhaps one could claim that it is the size which accounts for the incomparability.[16] But, this would of course also rule out that adding a person with a life not worth living to a population makes it worse, so it would not imply the strong version of the asymmetry.

We should therefore conclude that interpreting the strong version of the asymmetry in terms of value incomparability or indeterminacy does not succeed in explaining the strong asymmetry. The asymmetry, on these views, amounts to making a distinction where there does not seem to be any relevant difference. What seems to be *ad hoc* about the asymmetry is not explained by the axiological approach. If the asymmetry is to be defended on axiological grounds then it will be by simply claiming that adding people with lives not worth living makes a state of affairs worse, but adding people with lives worth living does not make it better. But, the asymmetry has then not been explained, merely reformulated, and is reduced to a brute axiological fact.

## 7.3 The Person-Affecting Principle

In an attempt to get around the problems mentioned above one might claim that the addition of a person to a population does not make the population better or worse, other things being equal, because we have not taken the value *for* people into account. It is not worse *for* anyone if we do not add a person with a life worth living to a population, and therefore it cannot be worse not to add a person who would have a life worth living to a population. Perhaps the intuition of neutrality, and the asymmetry, could be saved by saying that whether one population is better (or worse) than another is constrained by whether either is better (or worse) *for* someone.

The idea that goodness depends on what is good for people is a popular idea. However, there are many ways to describe the relation between good

---

$F$ than the other. But, for other concepts there is a version of $x$, $x_+$, which is clearly more $F$ than $y$ than and a version of $x$, $x_-$, which is clearly less $F$ than $y$. For example, if it is vague whether $x$ is *hairier* than $y$, then by adding enough hair to $x$ we get something which is hairier than $y$. Likewise, by removing hair from $x$ we get something which is less hairy than $y$. The indeterminacy of betterness can not, if it is to explain the asymmetry, work in this way. We can make a population (determinately) worse by adding people with lives not worth living, but we cannot make a population (determinately) better by adding people with lives worth living.

[16] See for example Brown (2011).

and good for. In its simplest form the view can be formulated in the following way:

> *Person-Affecting Principle*: if *A* is better (worse) than *B* then *A* is better (worse) than *B* for some person *p*.[17]

This principle, as it stands, does not imply anything in particular regarding the asymmetry because it is not clear how to interpret it in cases where *p* only exists in one of *A* and *B*. A common view is that *A* is better (worse) than *B* for *p* only if *p* exists in *both A* and *B*.[18] But, this interpretation of the Person-Affecting Principle clearly rules out the asymmetry. While it follows from this interpretation of the principle that adding a person with a life worth living cannot make a population better, because existing with a life worth living is not better for the person than non-existence, it also follows for the same reason that adding a person with a life *not* worth living cannot make a population worse. This version of the Person-Affecting Principle would therefore fails to imply the asymmetry.

Note also that this interpretation of the Person-Affecting Principle contradicts *Q* in certain same-number cases. For example, in a case like the young girl it would not be worse, according to this interpretation of the principle, for her child if she does not wait. It would not be worse *simpliciter* if she does not wait because it would not be worse *for* anyone.[19]

In order for the Person-Affecting Principle to have any hope of implying the asymmetry we must therefore assume that *A* can be better (worse) than *B* for *p* even though *p* does not exist in both. We would therefore have to assume that a life worth living is better for a person than non-existence and that a life not worth living is worse for a person than non-existence. While these are very controversial assumptions,[20] they will not be questioned here because even with these assumptions the principle fails to imply the asymmetry.

To see this, note first that because the Person-Affecting Principle only states a necessary condition it does not follow that adding a person with a life not worth living (i.e., negative well-being) to a population makes the population worse. A suggestion at this point is that we should appeal to the following principle:

> *Dominance*: If *A* is better (worse) than *B* for someone, and at least as good for everyone else, then *A* is better (worse) than *B*.

---

[17] Temkin (1993, p. 248) has dubbed this view "the Slogan". Temkin formulates it in the negative however: "[o]ne situation *cannot* be worse (or better) than another if there is *no one* for whom it *is* worse (or better)".

[18] See for example Broome (2004, p. 65).

[19] For further arguments against this interpretation of the Person-Affecting Principle, see Arrhenius (2009, pp. 295–6) and Broome (2004, pp. 135–6).

[20] See for example Holtug (2001), Roberts (2010), Johansson (2010) and Arrhenius & Rabinowicz (2010).

This principle seems plausible and is sometimes assumed to be a part of the Person-Affecting Principle.[21] It also follows from the Person-Affecting Principle and Dominance that adding a person with a life not worth living to a population makes the population worse, other things being equal, so if the Person-Affecting Principle is to imply the asymmetry then we seem to have good reason to accept Dominance. However, note that it also follows from the Person-Affecting Principle and Dominance that adding a person with a life worth living, i.e., positive well-being, to a population makes that population *better*, other things being equal. But, this contradicts the second half of the asymmetry. According to the asymmetry in its axiological interpretation, adding a person with a life worth living to a population does *not* make the population better, other things being equal.

The Person-Affecting Principle has been developed in various ways. Two of the most discussed versions are *actualism* and *necessitarianism*.[22] According to both these views it is a mistake to think that the Person-Affecting Principle should take the well-being of merely possible people into account. Rather, the principle should be more restricted in its scope. According to actualism only the well-being of those who have existed, do exist or will exist matter when it comes do determining the value of a population. According to this view, if $A$ is better than $B$ for $p$, then this is only relevant to the evaluations of $A$ and $B$ if $p$ is an *actual* person. Necessitarianism, on the other hand, is the view that the Person-Affecting Principle should be restricted to those who do exist necessarily relative to the alternatives. On this view, only the well-being of those who would exist irrespective of what we do matters to the value of a population.[23]

Both actualism and necessitarianism have been criticised at length elsewhere in the literature.[24] Necessitarianism is the least plausible of these two, especially since we want an axiological theory which is compatible with $Q$ and with the strong asymmetry. As should be obvious, necessitarianism is not compatible with $Q$ because no one exists necessarily relative to the alternatives in same-number cases. That is, if we can create $p$ with a well-being of 5 or $q$ with a well-being of 10 then creating $p$ is not worse than creating $q$ according to necessitarianism. Furthermore, necessitarianism does not imply the strong asymmetry because necessitarianism gives us no reason to think that adding people with lives not worth living to a population makes the population worse. Rather, it implies that adding contingent people, whether they

[21] See for example Temkin (1987, p. 166) and Arrhenius (2000, p. 118).

[22] See Singer (1993, p. 103) for a version of necessitarianism. Actualism has been defended by Jackson & Pargetter (1986), Bigelow & Pargetter (1988) and Parsons (2002).

[23] More modest versions of actualism and necessitarianism claim that we should discount, but not disregard completely, the well-being of non-actual or non-necessary people. However, these more modest versions only support weak asymmetry.

[24] For example in Bykvist (2006), Arrhenius (2000, 2003) and Roberts (2010).

would have lives worth living or not, would not make a difference to the value of a population.

When it comes to actualism there are mainly two objections. First, assuming some connection between value and norms, actualism implies that what you ought to do can depend on what you will do. This is unacceptable, it has been argued, since it rules out deliberation. You cannot know all the relevant facts pertaining to what you ought to do and still deliberate if what you in fact will do is one of the relevant facts. Oughts, it seems, are not like this.[25]

The second objection is that actualism implies dilemmas of an especially problematic kind. Normally, a dilemma is a situation where every alternative is wrong. Actualism, however, implies that there are situations where you cannot avoid doing wrong as in a normal dilemma but also that whatever you do there will be an alternative which is *right*.[26] Suppose, for example, that the only alternatives are to create either of two people, *a* or *b*, whose lives would be not worth living. Suppose we decide to create *a*. Since *a* is actual her well-being, the fact that her life is not worth living, counts against creating *a*. However, since *b* is not actual if we create *a* we can disregard *b*'s well-being. If other things are equal it would then be the case that *if* we create *a* then it is better to create *b*, and vice versa.

Most importantly, regardless of whether these objections are decisive against actualism, actualism does not imply the strong asymmetry since it gives us no reason to doubt that it is better to add people with lives worth living. If we add a person with a life worth living then her well-being counts in full and therefore makes the world better. For the actualist version of the Person-Affecting Principle to imply strong asymmetry one would have to deny that adding actual people with lives worth living makes a population better. But, this would not be to explain the asymmetry. It would be to reassert it as an unexplained axiological fact.

### 7.3.1 Variabilism

A more sophisticated defence of the asymmetry is Melinda Roberts' "variabilism". As we will see, Roberts' approach to the asymmetry shares certain features with the Person-Affecting Principle discussed above, though her approach differs in that she does not formulate her defence of the asymmetry in axiological terms.

Roberts' formulates variabilism in the following way:

[25] The claim that the normative status of an act cannot depend on whether the act is performed is sometimes referred to as the principle of normative invariance. See Carlson (1995, 2002). Whether foreknowledge "crowds out" deliberation is however a contested matter. Bykvist (2007*b*), for example, argues that there are examples where it seems plausible to say that an acts moral status *does* depend on whether or not it is performed. What is crucial with respect to deliberation is not this dependence as such but rather whether a theory which allows for such dependencies can still be action-guiding.

[26] See Bykvist (2006, p. 274-5).

> The loss incurred at a world where the person who incurs that loss does or will exist has *full moral significance* both for the purposes of evaluating the act that imposes that loss and for the purpose of evaluating any alternative act that avoids that loss, while a loss incurred by that very same person at a world where that person never exists at all has *no moral significance whatsoever*. (Roberts 2011, p. 356).[27]

A "loss" is then defined in terms of what is better for a person: a person $p$ suffers a loss in a world $w$ if and only if there is an accessible world $w'$ such that $w'$ is better than $w$ for $p$.[28]

Roberts' definition of loss raises the question whether it can be better (worse) for a person not to have existed. Roberts' view is that it can. On her view, $w'$ is better than $w$ for $p$ if and only if $p$ has more well-being in $w$ than in $w'$. If $p$ does not exist in one of these two worlds, say $w'$, then we should say that $p$ has zero well-being in $w'$.[29] This means that a person can suffer a loss in worlds where she does not exist if there is an accessible world where she has positive well-being. This also entails that a person with negative well-being suffers a loss if there is an accessible world where she does not exist. If $p$ does not exist in either of two worlds then they are equally good for $p$.

We can now see how variabilism purports to explain the two parts of the asymmetry. First, if a contingent person's life would be worse than non-existence then there is an accessible world, one where she does not exist, which is better for her. Therefore, she suffers a loss in the world where her life is not worth living and there is a reason against bringing her into existence. Regarding the second half of the asymmetry we can say that if a person would have a life worth living, were she to exist, then she suffers no morally significant loss by not being created. She suffers a loss but it is not a morally significant one.

One objection to variabilism is that it only takes morally significant *losses* into account and not morally significant *gains* (or benefits). Variabilism entails that the *loss* a person with a life worth living would suffer in worlds where she does not exist does not give us a reason to create her but so far variabil-

[27] See also Roberts (2010, ch. 2).

[28] In Roberts (2010) the definition reads "a person incurs a loss whenever agents (by act or omission) create less wellbeing for that person when agents could have created more wellbeing for that very same person" (Roberts 2010, p. 46). In Roberts (2011) she says that "to say that a person $p$ incurs a *loss* at a given world $w$ as a result of a given act $a$ is to say that there was still another world $w$' accessible to agents at the critical time such that their performance of an alternative act $a$' at $w$' is better for $p$ than their performance of $a$ at $w$ is" (Roberts 2011, p. 337). On both these formulations there is a reference to agents and their alternatives. However, whether a person incurs a loss, on Roberts' view, is fully determined by the better-for relations which hold between a possible world and the accessible alternatives. See Roberts (2011, p. 337).

[29] See Roberts (2011, p. 338).

ism does not entail anything about whether the fact that she has a life worth living counts in favour of creating her. Strictly speaking, it is compatible with variabilism to claim that a possible future person's happiness counts in favour of creating her. What variabilism rules out is that this person would suffer a morally significant loss by being left out of existence, but this is clearly a different question than whether the fact that she would have a life worth living were she to exist is a reason to make her exist.

In reply to this objection, Roberts has argued that we should understand benefits to simply be the converse of losses. She writes that

> gains are important on the same variable basis on which *losses* are important. More specifically: gains have moral significance, not when those gains are accrued at the world at which the person who accrues those gains *exist*, but rather when the *losses* those gains *avoid* on behalf of that person are incurred at worlds where *the person who incurs those losses exists*. (Roberts 2011, p. 365).

Since "better" and "worse" are interdefinable, we can define benefits by simply replacing "better" with "worse" in the definition of morally significant losses:

> $p$ enjoys a morally significant *gain* at a world $w$ iff there is an accessible world $w'$ which is *worse* for $p$ and $p$ exists in $w'$.

With this definition of morally significant gains we can then see how variabilism accounts for the second claim in the asymmetry. The gains a contingent future person would enjoy, were she to exist, are only morally significant if it would be worse for this person if she did not enjoy these gains *and* this person would exist if she did not enjoy these gains.

Taking gains to be the converse of losses comes with a fairly steep price however. One feature of Roberts' definition of a morally significant loss is that it is impossible for a person to suffer a morally significant loss in worlds where she does not exist. However, because Roberts takes morally significant gains to be the converse of morally significant losses, it follows that a person enjoys a significant gain only if she exists, not in the world where she enjoys the gain, but in an accessible world which is *worse* for her. This means, among other things, that we can bestow a morally significant gain to a person by *not* creating her.

Note that it is necessary to analyse gains in this way if variabilism is to imply the asymmetry. If one were to say that a person can only enjoy a morally significant gain in worlds where she exists then, contrary to what the asymmetry claims, that gain would be morally significant. Creating a person with a life worth living would in that case be to bestow a morally significant gain to that person because there is an alternative which is worse for her (non-existence).

Variabilism therefore requires gains to be analysed as the converse of losses in order for it to imply the asymmetry.

Why should we think that there is this difference between morally significant gains and losses? That is, why can morally significant losses only be suffered by a person in worlds where that person exists while morally significant gains can be enjoyed by people in worlds where they do not exist? To this question the only forthcoming answer seems to be that this difference between gains and losses is necessary in order for variabilism to imply the asymmetry.[30] But this is not a sufficient reason to consider the asymmetry saved, nor does it remove what at least appears to be a conflict between its two constituent claims. Rather, it shows that variabilism presupposes the very view, the asymmetry, which it purports to justify.

## 7.4 Summary

In this chapter I have argued that axiological approaches which are congenial with *Q* cannot explain the strong asymmetry. The problem with the asymmetry, from an axiological perspective, is that it has to be explained why adding a person with a life not worth living to a population makes it worse while adding a person with a life worth living to a population does not make the population better.

Both impersonal and person-affecting axiologies imply either that adding a person with a life worth living to a population makes the population better, or that adding a person with a life not worth living does not make a population worse. If an axiology is to be consistent with both claims then the asymmetry would have to be reduced to a brute axiological fact which does not allow for further explanation.

It might, at this point, be claimed that the failure of the axiological approaches is sufficient reason to abandon the strong version of the asymmetry and settle for a weaker version instead. This conclusion would be premature however. As I will argue in the next chapter, the asymmetry can be defended if we focus on the Harm Principle and an analogous Principle of Beneficence. The Harm Principle, I will argue, is not an explanatory fifth wheel. Rather, it does some serious work in explaining the asymmetry.

[30] Roberts seems to acknowledge this point. See Roberts (2011, p. 365).

# 8. The dual role of harm

In the previous chapter I argued that axiological approaches to population ethics fail to capture the intuition that there is an asymmetry in our obligations to future people. The strong version of this asymmetry consists of the following two claims:

> (I) If a contingent future person would have a life not worth living then this is a reason against creating that person.
> (II) If a contingent future person would have a life worth living then this is not a reason in favour of creating that person.

The asymmetry is a common view but it calls out for an explanation. As I have argued it is very difficult to formulate an axiology which explains the asymmetry and which is also consistent with the same-number quality claim, *Q*.

In this chapter I will consider whether a plausible defence of the asymmetry might be found if we focus on two normative principles, the Harm Principle and an analogous Principle of Beneficence. I will argue that the strong asymmetry can be explained by distinguishing between two kinds of reasons: requiring and (merely) compensating. According to the defence of the asymmetry which I will suggest benefits to contingent future people do provide a kind of reason, but this kind of reason is not sufficient to make it required to create a person with a life worth living. Harms to contingent future persons can however make it required not to create a person with a life worth living because harms provide reasons of the requiring kind. This defence of the asymmetry does not preserve (II) to the letter because benefits to contingent future persons *do* provide reasons in favour of creating them. However, I will argue that it preserves it spirit.

I will also argue that appealing to these two kinds of reasons involves assigning a "dual role" to harm. First, harms provide reasons as specified by the Harm Principle. Second, harm is a precondition for benefits to provide reasons of the same kind. I will argue that the distinction between two kinds of reasons is presupposed by two normative categories which are a pervasive part of common-sense morality, supererogation and options, and that the asymmetry can be defended as an instance of an option. I will argue that the most plausible ground for this option is the importance of autonomy. Grounding the option to create people with lives worth living involves assigning a special importance to harm which explains why harms play these two roles in procre-

ative decision and not benefits. Finally, I will consider some objections which could be raised against this defence of the asymmetry.

## 8.1 Harms and benefits

It seems easy enough to argue for the first half of the asymmetry by appealing to the Harm Principle. If a person would have a life not worth living then the bad things in that life outweigh the good things. Such a life therefore contain at least one negative well-being component. If one were to bring a person into existence with a life not worth living then there would be a state of affairs which makes that person's life go worse in the sense specified by (2*) in the Minimalist View (see chapter 4). One would also be responsible for that state of affairs because the state of affairs would not have obtained had one not created this person (see chapter 5). This means that bringing a person into existence with a life not worth living would be to harm her and we therefore have a reason against doing so.

The Harm Principle offers no guidance regarding the second half of the asymmetry. This is because the second half of the asymmetry is concerned with the *benefits* a future person would enjoy rather than the harms she would suffer. We therefore need to consider benefits and under what circumstances they provide reasons.[1]

The most straightforward approach would be to adopt a principle of beneficence analogous to the Harm Principle:

> *The Principle of Beneficence*: if an act would benefit someone then this is a reason in favour of performing that act.

We could also analyse "benefit" in a similar way as "harm" was analysed in previous chapters. By replacing "worse" with "better" in (2*) while leaving the other conditions as they are, we get the following analysis:

> *a benefits b* if and only if
> (1) *a* performs an act, $\phi$,
> (2*) *b* is in a state *S* such that *b*'s life is *better*, taking *S* into account, than *b*'s life not taking *S* into account.
> (3*) if $\phi$ had not been performed then *S* would not have obtained at all, or would have obtained in a different way which is salient in the circumstances.

[1] It might be claimed that the easiest way to defend the asymmetry would be to deny that benefits provide reasons under any circumstances. However, this view is very counter-intuitive. It suggests, for example, that the balance of reasons is almost always against creating people (because most lives include at least some harmful states). I discuss this kind of "anti-natalism" below.

It is easy to see however that the Principle of Beneficence together with this analysis of benefits would be incompatible with the second half of the asymmetry. A life worth living will include some benefits, i.e., some states which make the person's life better for her. According to the Principle of Beneficence these benefits would therefore be a reason to create her.[2] In order to save strong asymmetry we need a reason for thinking that we should disregard these benefits completely – the benefits a future person would enjoy give us no reason to create that person.

One such ground could be the versions of the Person-Affecting Principle discussed in the previous chapter. One could claim, for example, that we should disregard benefits to non-actual people (a version of actualism) or contingent people (a version of necessitarianism). These views are not much more plausible here than in the axiological context for two reasons. First, it would be *ad hoc* to apply this principle only to benefits and not to harms as well. Necessitarianism is the worst offender here. Applied to harm, necessitarianism implies that we should disregard harms to contingent future people. A necessitarian version of the Harm Principle would therefore not support (I). Second, the counter-examples against these constraints apply here as well. Actualism, for example, would still imply that there are situations where whatever we do there is an alternative which we have more reason to do. For example, suppose we can either create *a* or *b*, both of which would have lives not worth living. If we create *a* then *a* is actual and we have a reason against this action. However, since *b* is not actual there are no reasons against creating this person. Hence, there is an alternative which we have more reason to choose, other things being equal.

An alternative approach would be to restrict the Principle of Beneficence:

> *The Restricted Principle of Beneficence*: If an act would benefit someone then this is a reason in favour of performing that act *if and only if* the consequences of *not* performing the act would harm someone.[3]

Not all benefits are equal on this view. Those benefits that are not also preventions of harm do not provide reasons while those benefits that are also preventions of harm do. This principle is not satisfied when it comes to creating people who would have lives worth living because if the person had not

[2] Note that we do not need to enter the debate whether existence can be better than non-existence for a person in order to say that a person would be harmed, or benefited, by being created.

[3] For similar views, see Shiffrin (1999) and Benatar (2006). Note that this condition is not explicitly endorsed by Shiffrin or Benatar. They use slightly different terms but what follows is a plausible way of interpreting what they suggest. For example, Shiffrin formulates her defence of the asymmetry in Kantian terms and Benatar uses both personal and impersonal values. Neither of these are necessary to formulate this particular way of defending the asymmetry however.

been created then she would not have existed and therefore, would not have been harmed.[4]

The problem with this view is that it creates a presumption against procreation.[5] Suppose we could create a person who would live a very happy life. The Restricted Principle of Beneficence holds that the good things in this person's life do not provide reasons in favour of creating this person. However, it seems plausible to assume that this person would also suffer some bad things. They might not be very numerous, and the good things might be greater, but the claim that each future person we could create will suffer at least on occasion is difficult to deny. What this means is that it is reasonable to think that the Harm Principle is relevant: the person would be harmed were she to be created and the harms she would suffer provide reasons against creating her. The situation then is that there is no reason in favour of creating this person but some reason against.

By appealing to the Restricted Principle of Beneficence one is therefore committed to the view that there is a general presumption against creating new people. If we create a person then, because almost every life contains some bad things, we will harm her and this is supposed to be a reason against creating her. Furthermore, the benefits this person would enjoy do not provide us with a reason to create her. Defending the asymmetry on the basis of the Restricted Principle of Beneficence therefore implies anti-natalism: the balance of reasons is (almost) always against creating new people.[6] Anti-natalism is a view few are inclined to accept. Even though it is a consequence of this defence of the asymmetry, it seems more plausible to deny that there are no reasons in favour of creating happy people than to affirm that there is a presumption against procreation.

We can summarise the difficulty with strong asymmetry in the following way: on the one hand, if a person would benefit from being created then the benefits this person would enjoy are not a reason to create her. On the other hand, if the benefits do not provide reasons to create her then there does not seem to be anything to counterbalance the reasons against creating her stemming from harms. That is, we end up with anti-natalism. To defend the asymmetry *and* avoid anti-natalism it seems we will have to claim that the benefits a contingent future person would enjoy do not provide reasons to create her while at the same time claim that the same benefits must be able to provide reasons to create her.

It therefore appears to be the case the we cannot save the strong asymmetry while also rejecting anti-natalism. This, it might be thought, gives us excellent

[4] As I argued in chapter four, non-existence does not makes a person's life go worse, though it undermines that anything else makes a person's life go better or worse.

[5] This is argued by McMahan (2009, p. 53, 61). See also McMahan (1981, p. 120, fn. 28) and (1988, p. 35).

[6] Benatar endorses this view while Shiffrin merely says that procreation is not a "morally innocent endeavor" (Shiffrin 1999, p. 118).

reason to reject the strong version of the asymmetry and to retreat to defending the weak version instead. Recall that weak asymmetry grants that there are reasons in favour of creating people with lives worth living but that reasons to benefit carry less weight than reasons against harming. However, abandoning the strong version of the asymmetry would be premature. As I will argue below it is possible to save the spirit of the strong version of by making a distinction between different kinds of benefits.

## 8.2 Two kinds of reasons

What I suggest is that we can save the spirit of the strong asymmetry while rejecting anti-natalism. The difficulty with combining the asymmetry and the rejection of anti-natalism only arises with the assumption that if the benefits a contingent future person would enjoy do not provide reasons to create her then they cannot be relevant to the balance of reasons in procreative choices. What I suggest is that there are two kinds of normative reasons: *requiring* and (merely) *compensating* reasons. Requiring reasons count in favour (or against) acts and can ground a requirement to (not) perform the act should the balance of reasons turn out in favour of (against) the act. Compensating reasons on the other hand also count in favour (or against) acts but they cannot ground a requirement to (not) perform the act. They can merely compensate for other reasons when determining the balance of reasons. Requiring reasons, obviously, also have this feature. What distinguishes compensating reasons from requiring reasons is that the former *lack* something. Compensating reasons lack the possible "force" of making an act required. Should the balance of reasons turn out in favour of an act then the fact that there are merely compensating reasons in favour of performing the act is not sufficient for the act to be required. We should not, however, understand requiring reasons to be sufficient for a moral requirement. The notion of a requiring reason is rather the notion of a kind of reason that is necessary in order for an act to be required.

With this distinction between two kinds of reasons the strong asymmetry can be formulated in the following way: the harms a contingent future person would suffer provide requiring reasons while the benefits a contingent future person would enjoy provide (merely) compensating reasons. Creating a person with a life worth living is favoured by the balance of reasons, because the benefits outweigh the harms, but it is not required to create a person with a life worth living because the reasons that tip the balance are (merely) compensating. It is required however not to create a person with a life not worth living because these reasons are of the requiring kind.

If there is such a distinction between different kinds of reasons to be made then one could also argue that anti-natalism does not follow from the strong version of the asymmetry in the following way. When it comes to creating happy people, the bad things in that person's life provide reasons against do-

ing so. However, the good things (the benefits) provide (merely) compensating reasons in favour of creating the person. The balance of reasons is therefore *not* typically against creating people. Note that we need not say that benefits cannot provide requiring reasons. In some circumstances it is very plausible that they do. In order to save strong asymmetry we need only claim that benefits to contingent future persons are (merely) compensating.

It might be asked whether the distinction between two kinds of reasons preserves the strong asymmetry. As the asymmetry was formulated it consists partly in the claim that the well-being a person would have were she to exist is not a reason in favour of making her exist. It might be objected that this is not true on my suggestion because benefits to contingent future persons *do* provide reasons on my view, just not the same kind of reasons as harms. However, even though it does not preserve the strong asymmetry to the letter it does preserve its spirit. After all, the benefits a contingent future person would enjoy cannot make it required to create her, while harms to a contingent future person can make it required not to create her. This difference between harms and benefits to contingent future people seems to be at the core of the asymmetry. If we can preserve the intuition that harms and benefits to contingent future people are different from a moral point of view by distinguishing between two kinds of reasons then we would have preserved all that a defender of the asymmetry could ever have wanted.[7]

McMahan (2009) entertains a view along these lines. He argues that in order to save the strong version of the asymmetry we need to make a distinction between the "reason-giving function" and the "canceling function" of goods and bads. On his view, goods which only have a canceling function "do not count as reasons for causing the person to exist [since they do not have the reason-giving function]. But they do weigh against and cancel out corresponding bads that the person's life would contain" (McMahan 2009, p. 53). What the asymmetry presupposes, according to McMahan, is that some goods, such as benefits to future people, lack the reason-giving function but have the canceling function.

The way McMahan draws the distinction some goods, those that only have a canceling function, can be weighed against bads with a reason-giving function. A problem with this way of describing the distinction is that it is difficult to see how such "canceling goods" can be weighed against bads unless they do in fact provide reasons. For this reason it seems more fitting to say that all benefits provide reasons, they all have a "reason-giving function", but that some goods provide reasons which are merely compensating while others provide reasons which are requiring.

A similar distinction is also defended by Gert (2007).[8] Instead of saying that there are two *kinds* of reasons Gert argues that practical reasons in general

[7] This view is also not merely a version of the weak asymmetry because harms and benefits have the same weight, though they can differ in kind, according to this view.

[8] See also Gert (2000, 2003).

have two "dimensions" of normative strength: a *justifying* and a *requiring* dimension. A reason's justifying strength is the degree to which the reason justifies an act while a reason's requiring strength is the degree to which it requires an act. Gert argues that these two dimensions can come apart and that a reason can have great justifying strength but very little requiring strength.

The distinction between justifying and requiring normative strength could, if the distinction can be maintained, be used to defend the asymmetry in a similar way as McMahan's distinction. We can understand the justifying dimension of a reason as the reason's ability to make an act permissible. In order for an act to be required, however, it has to be favoured by reasons which have requiring strength. If it could be argued that the reasons supporting creating people with lives worth living has no requiring strength, but only justifying strength, then one could avoid anti-natalism by claiming that creating people with lives worth living is permissible ("justified") but not required.

Gert explicitly argues against the claim that there are two kinds of reasons because "any reason that can require me not to act on contrary reasons also justifies me in acting against those reasons, and therefore has both requiring and justifying strength" (Gert 2007, p. 542). Suppose, for example, that I would prevent future pain by going to the dentist and that the prevention of future pain requires me to do so. But, it seems plausible that I am also justified in going to the dentist, and that it is the same reason which does the requiring and the justifying. There is only one reason here, the prevention of future pain, and that this reason is what requires and justifies a visit to the dentist.

However, this example seems to work against Gert's view that there are two dimensions of normative strength. Gert assumes that reasons with requiring strength also have justifying strength but, if the two dimensions are independent of each other then it would be a coincidence that requiring strength coincides with justifying strength. That they do coincide is *not* a coincidence if we draw the distinction between two kinds of reasons as I did above. The difference between requiring and (merely) compensating reasons is that compensating reasons lack the ability to ground a requirement but they are alike in other respects. We do not have to say that there is one requiring and one compensating reason in the dentist-example. Rather, there is just one reason, the prevention of future pain, which is of the requiring kind.

On the other hand, if the two dimensions are not independent of each other, so that all reasons with some requiring strength have the same justifying strength, then it is unclear why we should not say that there are two kinds of reasons. One kind which has some requiring strength, and an equally strong justifying strength, and another kind of reason which has no requiring strength but some justifying strength. This possibility also suggests that the difference between my view and Gert's may not be that great. The main difference, it seems, is whether we should say that there are two kinds of reasons or just one kind with two "functions" or "strengths". This difference seems to be one

regarding the metaphysics of reasons and will not greatly affect the defence of the asymmetry which I will present below.

Gert also objects to the use of phrases such as "the balance of reasons".[9] Because there are two dimensions of normative strength, Gert argues, to talk about "the balance of reasons" is just a misleading metaphor. When weighing two reasons we cannot simply say that one outweighs, or is stronger, than the other in an unqualified way. On Gert's view we can talk about the balance of reasons with respect to requiring strength and the balance of reasons with respect to justifying strength, but there is no balance of reasons *period*.

Doing away with the balance of reasons makes Gert's way of drawing the distinction less attractive. Talk about "the balance of reasons" in an unqualified way seems to make sense, and drawing the distinction between two kinds of reasons has the advantage that it is less revisionary in this respect. As I will argue below, however, the balance of reasons may not be as *useful* as it is sometimes thought. But, we should not give up on the notion for this reason.

Gert's claim that it is pointless to talk about the balance of reasons raises the question what we should say, if there are indeed two kinds of reasons, about the relation between the balance of reasons and moral requirements. One objection which could be raised against making a distinction between two kinds of reasons is that it is incompatible with the following claim: if an act is favoured by the balance of reasons then it is morally required. Making a distinction between two kinds of reasons seems to be incompatible with this claim because the distinction amounts to a view where an act could be favoured by compensating reasons only. Such an act would not be morally required because compensating reasons cannot ground a moral requirement.

However, we should not be so eager to accept the view that if an act is favoured by the balance of reasons then it is morally required. For example, I may have excellent reasons to scratch an itch but that does not make me morally required to do so.[10]

While the worry just mentioned does not seem especially troubling, it might be claimed that a plausible requirement on the relation between reasons and duties is that if an act is favoured by the *moral reasons* then it is morally required. If compensating reasons are moral reasons then, the objection goes, the view above would not be compatible with this requirement. However, it is

[9] See Gert (2007, pp. 548–9).

[10] An interesting question, which I will here merely mention, is what to think of the reverse conditional: if an act is morally required then it is favoured by the balance of reasons. If we accept the distinction between compensating and requiring reasons then we would have to deny even this claim because it seems possible that the act which is favoured by the balance of reasons only has compensating reasons favouring it while an inferior alternative is supported by requiring reasons. If a plausible case of this kind could be construed then I should reject this claim as well. See however Stroud (1998) who argues that a restricted version of the reverse conditional is the only plausible version. A possible restricted version which would be compatible with the two kinds of reasons is the following: if an act is morally required then it is favoured by the balance of requiring reasons.

not entirely clear what is meant here by "moral reasons". One way to understand "moral reason" is to say that a reason in favour of $\phi$-ing counts as moral if and only if the reason could ground a moral requirement to $\phi$. On this interpretation the requirement seems trivial and it is quite clear that the distinction between compensating and requiring reasons is compatible with it. According to this view of "moral reasons", compensating reasons are not moral reasons at all.

An alternative way to understand "moral reasons" is to say that a reason counts as moral if and only if the reason is relevant to the moral status of an act (or a state of affairs). On this interpretation the distinction between requiring and compensating reasons would be inconsistent with the requirement.[11] However, interpretating moral reasons in this way makes the requirement less trivial. As I will argue below, two aspects of common sense morality, supererogation and options, are arguably inconsistent with this requirement. We should therefore not accept the alternative notion of moral reasons in so far as we think that supererogation and options make sense.[12]

A more serious objection is that the distinction between two kinds of reason is artificial and that there are no grounds, independent of the asymmetry, to accept it. Relying on this distinction is no more plausible than a defence of the strong asymmetry based on the axiological approach's "brute axiological facts". McMahan, for example, raises this worry and writes that introducing the two functions he mentions seems "strikingly ad hoc" (McMahan 2009, p. 54).

On closer examination one can say that the worry mainly concerns two things. First, the claim that some benefits count in one way under some circumstances and in another way in other circumstances seems *ad hoc*. It is not very plausible to claim that benefits never provide requiring reasons, so why do not all benefits provide the same kind of reason? Second, even if we accept the distinction between two kinds of reasons, why should we think that benefits to contingent future people provide reasons of the compensating kind? Also, why do harms to contingent future people provide requiring reasons and not compensating reasons?

In reply to the first worry I will argue that there are striking structural similarities between the two kinds of reasons relevant to the asymmetry (favouring and requiring) and reasons that figure in two normative categories: supererogation and options. These similarities support the conditional claim that

[11] There are certainly details to spell out here. It seems plausible however that compensating reasons are moral reasons in this sense because they are relevant to whether an act is permissible, as the asymmetry illustrates.

[12] This applies even to those who reject supererogation and options because they think that there are no acts which have these properties. That is, anyone who thinks that supererogation and options are *conceptually possible* should reject the suggestion that "moral reasons" should be understood as those reasons which are relevant to the moral evaluation of an act or a state of affairs.

if a moral theory includes options or supererogation then there exists the necessary conceptual space required by the asymmetry. In reply to the second worry I will suggest that benefits to contingent future persons provide (merely) compensating reasons because it is important to allow people to pursue their own aims, interests and projects as long as they do no harm. That people are so allowed, I will argue, is a way of respecting people's autonomy. What "grounds" the option to create people with lives worth living, and options in other circumstances, is the value of autonomy thus conceived.

Autonomy, understood as the freedom to pursue one's own aims, interests and projects as long as they do no harm, also answers why *harms* to contingent future persons provide requiring reasons and not (merely) compensating reasons. Avoiding harm is more important than benefiting because harms play this second role, besides providing reasons, of setting the limits of autonomy.

### 8.2.1 Supererogation

Supererogatory acts are usually characterised as acts that are "beyond the call of duty".[13] Typical examples are sacrificing one's life in order to save others', devoting all one's life and resources to help the less fortunate and so on. Such acts are not usually thought to be required by morality even though there are usually strong reasons in favour of them. Also, supererogatory acts are generally, if not always, praiseworthy.

A possible approach to supererogation is to maintain that while there are strong reasons in favour of such acts, there are also strong reasons against them. Consider the examples mentioned above. Sacrificing one's life to save others' is, after all, a great cost to the agent. It might be claimed that this cost to the agent should be accounted for in the balance of reasons. On this view, a supererogatory act is not favoured by the balance of reasons because the cost to the agent of performing the supererogatory act has to be factored in.

I mention this as a possible view, but it is not plausible as a view of supererogation. On this view, supererogatory acts would be those where the reasons provided by the cost to the agent and the reasons provided by benefits to others are perfectly matched. Otherwise, if the cost to the agent outweighed the benefits to others then the agent would be required not to perform the sacrifice. This is not how supererogation is typically thought to work. Rather, a supererogatory act is permissible, but not required. This view also fails to capture the idea that to perform a supererogatory act is to do more than is required. It is central to supererogation that the act is supported by the best reasons, all things considered, but that it is nevertheless not required.

Also, on this account of supererogation we would have to give the cost to the agent an implausible weight if typical examples of supererogation are not favoured by the balance of reasons. For example, suppose that Blue can save

[13] See Urmson (1958) and Heyd (1982).

the lives of ten strangers by throwing himself on a bomb, thereby shielding the others from the blast, and that throwing himself on the bomb is supererogatory. There is then a reason in favour of making the sacrifice (it will save ten lives) and, according to this view of supererogation, a reason against making the sacrifice (it will cost Blue his life). However, note that if Blue does not save the ten strangers then he is, in effect, saving himself. Blue's choice is therefore a choice between saving one, which happens to be himself, and saving ten strangers. In order for this to be a case where there is a tie in the balance of reasons it seems that we would have to say that Blue is allowed to give saving his own life an implausibly greater weight than saving a stranger's life.

It could of course be objected that this is *not* an example of supererogation. That is, it could be claimed that Blue is actually required to makes the sacrifice. While this is a possible view it is simply to deny the premise that this is an example of supererogation. It is therefore not a reply which is friendly to supererogation in general. After all, sacrificing oneself to save others is supposed to be a typical example of supererogation.

Supererogation, if it exists, therefore seems to assume that there is a distinction among reasons similar to the distinction between requiring and compensating reasons. Supererogatory acts are favoured by the balance of reasons but they are not required. The reasons in favour of performing supererogatory acts are therefore unable to ground a moral requirement. They can however be weighed against other reasons. It therefore seems plausible to say that supererogation presupposes the existence of compensating reasons. In reply to the first worry mentioned above it could then be argued that making a distinction between two kinds of reasons is not *ad hoc* because the same distinction is presupposed by supererogation.

The second worry was that even if there is a distinction between two kinds of reasons to be made, it still has to be shown that benefits to contingent future people provide (merely) compensating reasons and not requiring requiring. Here appealing to supererogation will not be of much help because there are significant differences between creating people with lives worth living and supererogation. For one thing, supererogatory acts are praiseworthy but this is not true of procreation. Of course, most people consider having children a good thing but they hardly think themselves heroic or saintly in having children, and rightly so. This difference suggests that while supererogation and the asymmetry seem to assume that there are two kinds of reasons, the latter can not plausibly be thought of as an instance of the former.[14]

[14] Heyd (1994, p. 115) notes and dismisses the possibility of classifying the asymmetry as an instance of supererogation for precisely this reason. It has been argued by Chisholm (1963) that what characterises supererogatory acts is that while it would be good if they were performed it would not be bad if they were not performed. If one adopts Chisholm's suggestion then it would be plausible to say that creating people with lives worth living is supererogatory. However, as the connection with praiseworthiness indicates, Chisholm's criteria for supererogation

We should therefore conclude that while supererogation presupposes two kinds of reasons, requiring and compensating reasons, the asymmetry can not plausibly be viewed as an instance of supererogation. Let us turn to options and see whether they make a better case for the asymmetry.

### 8.2.2 Options

A more plausible approach to the asymmetry is to characterise creating people with lives worth living as "optional". To say that agents have a moral option to $\phi$ is to say that it is permissible but not required for agents to $\phi$, even if $\phi$-ing is not supported by the balance of reasons. Typical examples of options are giving priority to the well-being of one's near and dear over the well-being of strangers or to give priority to one's personal projects over others'. It is on this view permissible to, say, save your drowning child rather than a stranger's, or to work on your own dissertation rather than helping a colleague, even if saving the stranger's child or helping your colleague is supported by the balance of reasons.

One attempt to account for the permissibility of saving your child rather than a stranger's is to say that the personal tie, the fact that it is *your* child, provides *you* with a reason which is not applicable to anyone else. On this view, you have an agent-relative reason which only applies to you. It is permissible to save your child because even if the benefits of saving the stranger's child are greater than the benefits of saving your child there is an agent-relative reason, the personal tie, which counterbalances these. Options, on this view, are ties in the balance of reasons.

However, reducing options to ties in the balance of reasons does not seem plausible because it is characteristic of options that they are "suboptimal" in the sense that they are not favoured by the balance of reasons yet still permissible.[15] This characterisation of options is also a pervasive part of common sense morality. Consider certain counter-examples to act utilitarianism which are designed to show that this theory is too demanding. Such counter-examples start from the observation that in our everyday lives we usually indulge in small enjoyments even though our time and money could be better spent elsewhere. For example, it is morally permitted to occasionally go to the cinema. However, it is quite plausible that more good could be done by doing something else, such as giving the money spent on the ticket to charity. Still, we think that we are not doing anything wrong by occasionally visiting the cinema. However it is hard to deny that on every particular visit we do in fact have stronger moral reasons to give the money to charity instead. What this suggests is that options presuppose that the reasons which favour giving the money to charity are not sufficient to make it required to give the money

---

are, at best, necessary and not sufficient. See also Heyd (1982, pp. 113–120) for a critique of Chisholm's analysis of supererogation.

[15] See Kagan (1989, pp. 3–4, 75–6) and Scheffler (2003, pp. 22–3).

to charity. That is, options seem to presuppose a *different kind* of reason. The balance of reasons are in favour of giving money to charity but the reasons that tip the balance are not sufficient to make it required.

Options, just as supererogation, seem to presuppose a kind of reason that is relevant to the balance of reasons but which are non-requiring. The main difference between supererogation and the asymmetry is that supererogatory acts are generally considered praiseworthy while creating new people is not. Options, on the other hand, do not have this connection to praiseworthiness and it is therefore more plausible to say that the asymmetry is an instance of an option.[16] We have not shown, of course, that there are options nor that creating people with lives worth living is optional. What the structural similarities between options and the asymmetry suggests is that if there are options, then the asymmetry could be defended as an instance of an option.

#### 8.2.2.1 Options and autonomy

Even if there are options it still has to be shown that creating people with lives worth living is optional. In order to establish that an act is optional one has to show that there is some special consideration which makes it plausible that a suboptimal act is permitted. What is lacking, in other words, is a plausible "ground" for the option.

One view is that what grounds an option has to do with the cost to the agent of performing the optimal act. If performing the act which is favoured by the balance of reasons would require that the agent sacrificed too much then the agent is allowed to perform some other act which avoids this sacrifice. More precisely, the suggestion can be formulated in the following way:

> *The Appeal to Cost*: if the cost to the agent of performing an act which is favoured by the balance of reasons is very great then the agent is allowed to perform an inferior alternative, provided that the alternative is not very inferior.[17]

There are a number of problems with this view as a ground for procreation as an option. First, grounding an option to create people with lives worth living

[16] One possible view of options is that they correspond to what Driver (1992) calls "the suberogatory". Suberogatory acts are "acts that we ought not to do, but which are not forbidden" (Driver 1992, p. 291). The intuition which Driver attempts to capture by this concept is that there seems to be acts which we disapprove of but which we nevertheless think are morally acceptable. It seems to me that while some optional acts may be suberogatory, it is not true of all options. Options, as I will understand them, have no necessary connection to praise, blame, approval or disapproval.

[17] See for example Scheffler (2003, p. 20), though Scheffler's version of the appeal to cost is a bit more complicated. He suggests that it would be permissible for an agent to "promote the non-optimal outcome of his choosing, provided only that the degree of its inferiority to each of the superior outcomes he could instead promote in no case exceeded, by more than the specified proportion, the degree of sacrifice necessary for him to promote the superior outcome". See also Kagan (1989, ch. 7).

based on the cost to the agent would be too strong because it implies that it would be optional to create people in cases where other people would benefit from the new person's existence. The asymmetry is silent with respect to *this* claim. Second, it is doubtful, at best, whether the costs of procreation are sufficient for typical cases of procreation to count as an option. It is a common view, at least, that raising a family is not a burden but a privilege. This suggests that the cost of procreation is typically not significantly large in order for it to be optional. Finally, it is unclear why there would be two kinds of reasons given the cost-approach and why benefits to contingent future people would only provide compensating reasons. Rather, according to the Appeal to Cost an agent could be permitted to harm contingent future people if the cost of not harming them would be too costly to the agent. This suggests that what the Appeal to Cost amounts to is not that there is a difference between benefiting contingent future people and harming them, but rather that agents are allowed to give a disproportionate weight to their self-interest.[18] The Appeal to Cost therefore fails to establish the asymmetry as an option.

An alternative view is to say that what grounds options is that people should be allowed to pursue projects which are important to them, even if pursuing these projects would be to perform sub-optimal acts. For example, Nagel (1986, pp. 166–70) argues that there are reasons of "autonomy" which stems "from the desires, projects, commitments, and personal ties of the individual agent, all of which give him reason to act in the pursuit of ends that are his own" (Nagel 1986, p. 165). Nagel goes on to describe the value of these personal projects and commitments in the following way:

> Most things we pursue [...] are optional. Their value to us depends on our individual aims, projects, and concerns, including particular concerns for other people that reflect our relations with them; they acquire value only because of the interest we develop in them and the place this gives them in our lives, rather than evoking interest because of their value. (Nagel 1986, p. 168).

It is clear that on Nagel's view it is important to allow people to pursue their own aims, projects and concerns, though he is quick to add that "[t]he crucial question is how far the authority of each individual runs in determining the objective value of the satisfaction of his own desires and preferences" (ibid.). Autonomy is not the freedom to do whatever one wants but to pursue one's aims, projects and concerns within certain limits.[19]

[18] The Appeal to Cost seems very similar to what Sidgwick called "the dualism of practical reason". Sidgwick's view on the matter was that "in the rarer cases of a recognised conflict between self-interest and duty, practical reason, being divided against itself, would cease to be a motive on either side; the conflict would have to be decided by the comparative preponderance of one or other of two groups of non-rational impulses" (Sidgwick 1981, p. 508).

[19] For Nagel, the importance of autonomy appears to be connected to his claim that some things, though not necessarily all, are valuable because people take on certain attitudes towards them. There are of course other views on what explains the value of autonomy. For example, J. S. Mill argues in *On Liberty* that "individuality" is an important component of a person's well-being.

A problem here is that it is not clear that Nagel's view supports the existence of options. Rather, his view is that a person's projects provide agent-relative reasons whose strength is not the same as the agent-neutral reasons these projects provide. What Nagel's reasons of autonomy amounts to is that people are allowed to give a disproportionate weight to their own interests but it is not clear that people are allowed to ever perform an act which is not favoured by the balance of reasons.[20]

However, we should not be so quick to dismiss the idea that autonomy, in the sense that people are allowed to pursue their own aims and commitments within certain limits, can ground options. This idea is independent of Nagel's claims about agent-relative and agent-neutral reasons and it is worth exploring further without relying on agent-neutral and agent-relative reasons.

An alternative to Nagel's view about the importance of autonomy is to focus on the idea that the people are allowed to pursue their own aims within certain limits. In order for autonomy to ground an option to procreate however, we have to spell out what these limits are in a way which makes it plausible that the choice not to create a person who would have a life worth living are within these limits.

A view about what the limits to autonomy are is to say that people are allowed to pursue their own priorities within the limits of the Harm Principle: *only when an inferior alternative would do no harm* are we allowed to choose it. Options based on autonomy, on this view, amounts to the freedom to pursue one's own aims provided that doing so does no harm.[21]

Note that this way of understanding autonomy is not to say that there are reasons to pursue autonomy, nor that there are reasons "of autonomy" which weigh for or against different alternatives in a situation. Appealing to autonomy is *not* to say that there is an additional reason based on autonomy in favour of an otherwise sub-optimal alternative. Rather, the claim that people's

---

On Mill's view, the importance of being allowed to shape one's life after one's own priorities is important because it allows for a better life. See also Raz (1988, chs. 14–15).

A similar view is suggested by Williams (1984) who claims that people are permitted to pursue their own interests because doing otherwise would be "to alienate him in a real sense from his actions and the source of his action in his own convictions [...] It is thus, in the most literal sense, an attack on his integrity" (Williams 1984, pp. 116–7).

[20] That Nagel's view is not compatible with options has also been noted by Hurley (1995, pp. 167-8). Hurley also argues that Scheffler's account of options, based on the Appeal to Cost, does not allow for options (pp. 170–1).

[21] Scheffler (2003, pp. 182–4) suggests a "no-harm" prerogative which "would not give agents unqualified permission to devote proportionately greater weight to their own interests than to the interests of other people. Rather, it would only permit agents to do this provided they did not harm others in pursuit of their non-optimal ends" (Scheffler 2003, p. 183). See also Kagan's (1989, p. 188) criticism of a similar view which he calls "neo-moderete". According to the neo-moderete, it is permissible to pursue one's own interest if doing so does no harm. But, one is never permitted to perform acts which are sub-optimal and which do harm. Kagan's criticism is mainly targeted against a neo-moderete view formulated in terms of the Appeal to Cost however. His main objection to the kind of view sketched here is that it relies on a dubious act-omission distinction. I fail to see however how my view relies on that distinction.

autonomy should be respected is a normative claim about when an act is permissible:

> *The Principle of Permissibility*: an act is permissible if (i) it is favoured by the balance of reasons *or* (ii) it does no harm.

This principle states two conditions which are sufficient for an act to be permissible. The first clause is necessary because respecting people's autonomy does not imply that it is not *permitted* to do what is favoured by the balance of reasons. The principle is also compatible with it being required, in a specific situation, to perform the act which is favoured by the balance of reasons. If there is no alternative which does no harm then the only permissible act is the one which is favoured by the balance of reasons. The second clause captures the importance of autonomy. It is always permitted to perform an act which does no harm, so people are allowed to pursue whatever aims and projects they like as long as they do no harm.

Because the Principle of Permissibility only states two sufficient conditions for permissibility it leaves it open whether an act could be permissible for other reasons than (i) and (ii). For example, it might be claimed that people should be allowed to perform acts which do no *serious* harm. Extending the principle in this way might seem especially attractive considering that the analysis of harm which I have argued for is very wide. It might also be objected that the principle should allow people to harm themselves if it is to capture the importance of autonomy. For our present purposes we do not have to decide on these possible extensions of the Principle of Permissibility because both are compatible with (ii) being sufficient for permissibility.

The Principle of Permissibility is also congenial with the claim that harms always provide requiring reasons while benefits sometimes provide only compensating reasons. A way to understand the Principle of Permissibility is as assigning a second role to harm. On the one hand, harms provide requiring reasons. This is the content of the Harm Principle and explains why, for example, there is a reason against creating people with lives not worth living. On the other hand, harm also plays the role of being a precondition for benefits to provide requiring reasons by setting the limit on when a person is allowed to pursue her own aims over the general good. Because a person is free to pursue her own aims as long as she does no harm, we can say that a benefit provides a requiring reason only if the benefit prevents harm. This explains why benefits to contingent future people do not provide requiring reasons but only compensating reasons. The benefits a contingent future person would enjoy were she to be created only provide compensating reasons because no one would have been harmed had the person not been created.[22]

[22] Are there other examples where people would be allowed to perform sub-optimal acts because of autonomy, given that the analysis of harm is so wide? An example of a kind of case where autonomy could be relevant is a kind of case discussed by Bykvist (2006). Suppose your

Assigning a dual role to harm, and not to benefits, amounts to an asymmetry between harms and benefits of sorts. It could be objected that this asymmetry between harms and benefits is just as problematic as the original asymmetry.

It should be noted, first, that defending the asymmetry as an instance of an option, grounded in the Principle of Permissibility, is not *ad hoc*. On my view the asymmetry is defended by appealing to other considerations, options and autonomy, whose plausibility is independent of asymmetry. As I argued in the previous chapter, this is not typically the case for other ways of defending the asymmetry. In the case of "variabilism" for example, that view only implies the asymmetry if we adopt a counter-intuitive definition of benefits and the only reason for adopting this definition was that it was necessary for the theory to imply the asymmetry.

Second, assigning a dual role to harm captures the intuition that harms are more important morally than benefits. Harms are not more important in the sense that they typically provide stronger reasons. On the view just sketched we should give an equal weight to harms and benefits of equal sizes. Rather, harms are more important than benefits because all harms, regardless of the circumstances, provide requiring reasons. That harm actually plays this role is a normative claim which could be disputed. But, it is not a restatement of the asymmetry. It is rather a way of showing that the asymmetry can be defended as an instance of a more general phenomena, options, which is a pervasive part of common-sense morality.

The possibility of explaining the asymmetry by appealing to autonomy has also been suggested by Arrhenius (2013). Arrhenius argues that we should reject an axiological version of the asymmetry and endorse the claim that a contingent future person's well-being can be a reason to create her, but that

> [i]t doesn't follow, however, that these reasons are decisive in the sense that they can in themselves give rise to a moral obligation to procreate. [...] one can consistently hold the view that an addition of people with positive welfare might make an outcome better, other things being equal, but deny any obligation to procreate since one can appeal to other values or deontological considerations such as parental autonomy. (Arrhenius 2013, p. 222).

On Arrhenius' view, to say that the reason to procreate is not decisive is to say that such reasons are not enough, they are never "decisive in themselves" (Arrhenius 2013, p. 222) for there to be an obligation to procreate. The notion

---

well-being is determined by your preference-satisfaction and that you can choose between two careers, *A* or *B*. If you choose *A* you will come to prefer *A* and if you choose *B* you will come to prefer *B*. Assume also that your *A*-preferences, those you would have were you to choose *A*, would be stronger than your *B*-preferences. At the time of choosing you are indifferent between the two. In such a case you would not do any harm whatever you choose because your preferences are dependent on your choice. However, because your *A*-preferences would be stronger than your *B*-preferences, you would be benefited more by choosing *A*. Appealing to autonomy allows you in cases like this to choose either *A* or *B*.

of a reason not being decisive is very similar to my claim that the benefits a contingent future person would enjoy can not provide requiring reasons.

A problem for Arrhenius' view is that, for the appeal to autonomy to save the asymmetry, it has to be assumed that autonomy can make the reason to create a person with a life worth living non-decisive but, autonomy can not make the reason not to create a person with a life not worth living non-decisive. Why, for example, could not a couple who decides to have a child with a life not worth living defend this decision by appealing to their parental autonomy? A possible way for Arrhenius to reply to this objection would be to accept my claim above that there are limits to people's autonomy. The limits I have suggested to individuals' autonomy include that a person is only free to choose an inferior alternative if the inferior alternative would not do harm. A couple attempting to defend their decision to have a child with a life not worth living will therefore have no business doing so on the grounds of autonomy.

To summarise, I have argued that the asymmetry can be explained in a unified way by appealing to harm, autonomy and the distinction between two kinds of reasons. Important to this defence is the claim that harm plays a dual role. First, harm is connected with reasons as specified by the Harm Principle. This explains why there is a reason against creating a person with a life not worth living. Second, harm also plays the role of being a precondition for benefits to provide requiring reasons. The benefits a contingent future person would enjoy, were she to be created, are not also preventions of harm and do therefore provide (merely) compensating reasons, not requiring reasons. Benefits to contingent future persons can therefore only *compensate* for harm but they cannot make it *required* to create a person with a life worth living. This second role of harm is given by the importance of autonomy: benefits which are not preventions of harm only provide compensating reasons because people are allowed to shape their lives in accordance with their own priorities as long as they do no harm.

## 8.3 The asymmetry and the non-identity problem

One of the main tasks of this and the previous chapter has been to argue that a principle like *Q* does not render the Harm Principle redundant. However, the distinction between requiring and compensating reasons, and the claim that benefits provide requiring reasons only if they are also preventions of harm, might seem to be inconsistent with *Q*. *Q* entails that if one can create either of two persons who would have lives worth living then it would be better to create the person with more well-being. The Principle of Permissibility, on the other hand, implies that it permissible to create the person with less well-being if doing so does no harm. It might then be objected that the defence of the asymmetry suggested above forces us to give up too much.

Here it is important to note that *Q* is an axiological claim while I have been concerned with reasons. My view *is* consistent with *Q* in so far as I would agree that creating the person with more well-being is favoured by the balance of reasons and, in a sense, better. What one would have to deny, if we are to accept my defence of the asymmetry and *Q* is that one is always required to perform the action which is favoured by the balance of reasons (or best). This claim is inconsistent with the Principle of Permissibility and the distinction between compensating and requiring reasons as well as the existence of options as I have characterised them. But, it would not be inconsistent to endorse my view and *Q*.

Even if the defence of the asymmetry suggested above is strictly speaking consistent with *Q* it might be objected that this defence has counter-intuitive consequences regarding the non-identity problem. As was just noted, if we could create either of two persons with lives worth living then the fact that they would have lives worth living cannot make it required to create either of them, even if one would be better off than the other. It might also be objected that the view suggested above also has counter-intuitive implications in same-people cases. For example, it could be permissible on my view to create a person even when it is possible to create *the same* person but with more well-being. Some people find this counter-intuitive.

However, it should be remembered that it is only permissible on my view to perform a sub-optimal act if it does no harm. In order for it to be permissible to create a person with less well-being, whether one could create the same person with more well-being or not, we have to assume that creating the person with less well-being would be completely harm-less. Such a life is certainly strange and unfamiliar to us considering that my analysis of harm is very wide. What we are supposed to imagine is a case where there is nothing in the person's life which makes the life go worse. Appealing to intuitions about what we should do in such cases is therefore a bit sketchy. If there really is no harm involved in creating the person with less well-being, then it is unclear what it is that it so objectionable about it and why we are required to create the person with more well-being.

It might be wondered what kind of consideration would make it *allowed* to create the person with less well-being. I have suggested that the further consideration is the value of autonomy. On this view, people are allowed to pursue their own interests as long as they do no harm because this is at least a part of what it means to respect people's autonomy. In a case where one could create either of two people without doing harm it is therefore permissible to create either. It might of course be questioned whether autonomy matters in anything like the way I have suggested, and whether harm is that central to autonomy. But, these are further questions which take us beyond simple examples like the non-identity problem. Simply appealing to the non-identity problem as a

counter-example to the defence of the asymmetry sketched above is therefore not a very fruitful strategy.[23]

A number of authors have argued that there is a problem with the asymmetry if we compare it with the non-identity problem.[24] McMahan (2013, pp. 25–26), for example, considers the following cases:

Case 1.
$A = (10, -)$
$B = (-, 6)$
Case 2.
$C = (10, -)$
$D = (-, -)$

In these cases, a positive number represents that a particular person exists with a certain amount of well-being and a dash represents that a particular person does not exist. Case 1, for example, represents a choice between creating either of two people who would have lives worth living, but the *A*-person would be better off than the *B*-person.

When comparing these two cases, McMahan fails to find a morally relevant difference between *A* and *C*. It also seems plausible, according to McMahan, that *B* is worse than *A*. Furthermore, according to the asymmetry we should say that *D* is not worse than *C*. But, if that is the case then we should also say, according to McMahan, that *B* is worse than *D*.[25] But that is clearly false. It is not worse to create a person with a life worth living than to not create anyone.

However, the reply which was made above applies to McMahan's argument as well. There is a relevant difference on my view between the two cases, namely that in Case 2 there is an alternative (*D*) which does no harm. On the view suggested in this chapter, this feature is relevant to the evaluation of Case 2. Furthermore, my view is consistent with *B* being better, or favoured by the balance of reasons, when compared with *D*. What we can not say, on my view, is that *B* is required when *D* is an alternative.

Still, it has been argued that there is a problem for the asymmetry here. Bradley (2013) argues that the relative strength of a reason in favour of or

[23] It might also be questioned whether autonomy is relevant to all decisions which affect future people. Consider for example a "social planner" who has to decide on which policy to implement and that the policy will affect the number of people who will exist in the future. By choosing one policy, the planner can effectively "create" future people. Autonomy seems to be less relevant from this perspective than from a personal perspective. One important difference is that it is not clear that there are any personal aims at stake for the planner at all. For this kind of impersonal planner, therefore, it might only be permissible to choose the policy which is favoured by the balance of reasons.

[24] See Rachels (1998, p. 95, 103) and Belshaw (2003).

[25] If *A* and *C* are equally good, and *B* is worse than *A*, then *B* is also worse than *C*. If *C* and *D* are equally good, as McMahan thinks a defence of the asymmetry requires, then *B* is also worse than *D*.

against an act is not affected if we add further alternatives to that act.[26] The intuition behind this principle is easy enough to see. For example, if *a*'s alternatives are to torture *b* severely or not at all, then the reason against torturing *b* severely remains just as strong even when we add the alternative to torture *b* slightly. Bradley then argues that several attempts to save the asymmetry violate this principle and that the asymmetry should therefore be rejected.

Bradley's objection is similar to McMahan's. In Case 1 there is a reason in favour of doing *B* and, according to Bradley's principle, the strength of this reason should not be affected by whether *D* is an alternative or not. We should therefore say that there is more reason to do *B* than *D*, and therefore also more reason to do *C* than *D*.

Note however that all this is compatible with my view. On my view the strength of the reason in favour of *B* is unaffected by the addition of *D*. What is not unaffected however is the "quality" of the reason: the reason goes from being requiring to merely favouring because we add an option which would do no harm.

McMahan's and Bradley's objections do however highlight an important aspect of the way I have defended the asymmetry, namely the relation between reasons and duties. Because there are two kinds of reasons we cannot say that, necessarily, a person ought to do what she has most reason to do. I have suggested that this is not as counter-intuitive as it may seem by arguing that common-sense morality includes two normative categories, supererogation and options, which assume that people are sometimes allowed to perform acts which are not favoured by the balance of reasons. However, this means that the defence of the asymmetry suggested here involves a significant normative commitment.

Are these normative commitments reasonable? The final objection which I will consider is that my view rules out otherwise plausible claims about the ethics of procreation. Savulescu (2001), for example, defends the following principle:

> *Procreative Beneficence*: couples (or single reproducers) should select the child, of the possible children they could have, who is expected to have the best life, or at least as good a life as the others, based on the relevant, available information. (Savulescu 2001, p. 415).[27]

This principle is a clear statement of the view that we ought to have the best child and might therefore seem to conflict with my defence of the asymmetry. After all, on my view it is only required to have the best child if there is

[26] This principle is a version of the "independence of irrelevant alternatives". For a discussion of this principle and population ethics see, Arrhenius (2009).

[27] See also Kahane & Savulescu (2008). However, the version of Procreative Beneficence Kahane & Savulescu (2008) argue for is significantly weaker than the one stated here. According to their view, there is a significant moral reason to have the best child in procreative decisions but this reason is not necessarily decisive.

no alternative which does no harm. Couples would therefore, on my view, be permitted to have a child which would not be expected to live the best possible life when having this child does no harm.

In defence of my view there are a number of things to be said. First, we should bear in mind that benefits are not the only things that matter on my view and that my view implies a modified version of Procreative Beneficence. It *is* required to have the child with the better life if that child would be harmed less than the other child, for example. Second, my view is compatible with *Q* and we can therefore say that it would be better if couples (or single reproducers) selected the child with the best chances of living the better life. We can also say that having the child with the best life is favoured by the balance of reasons. We can even say that couples ought to have the opportunity to select the best child and that any relevant information in order to select the best child should be available to them. What we *cannot* say is that they ought, or are morally required, to have the best child. This, it seems to me, is not very counter-intuitive.

## 8.4 Summary

I have argued that the spirit of the strong asymmetry can be saved by making a distinction between two kinds of reasons: requiring and (merely) compensating. Harms, I suggested, always provide requiring reasons while benefits to contingent future people only provide compensating reasons. The spirit of the strong asymmetry can then be explained because while the fact that person would have a life not worth living is a requiring reason not to create that person, the benefits a contingent future person would enjoy can only compensate for reasons against creating this person. Creating people with lives worth living is therefore favoured by the balance of reasons but not required by morality.

This distinction between two kinds of reasons is presupposed by two normative categories: supererogation and options. Of these two, the asymmetry is clearly most plausible as an instance of an option because supererogatory acts are always praiseworthy. These structural similarities suggest that if there are options or supererogation then there exists the necessary conceptual space necessary in order to defend the asymmetry.

However, this similarity with options does not show that creating people with lives worth living is optional. I argued that one way of grounding such an options, the Appeal to Cost, does not succeed. Instead, I suggested that we should appeal to a Principle of Permissibility:

> *The Principle of Permissibility*: an act is permissible if (i) it is favoured by the balance of reasons *or* (ii) it does no harm.

This normative claim, I suggested, captures the importance of autonomy in the sense that it allows people to pursue their own, sub-optimal, aims as long as they do no harm.

Appealing to the Principle of Permissibility is also congenial with the distinction between two kinds of reasons. A way to understand the principle is as ascribing a second normative to harm in addition to the role specified by the Harm Principle. According to the Harm Principle, harms provide reasons. The second role of harm is that a benefit provides a requiring reason only if it also prevents harm. This role explains why it is permissible, for example, not to create a person who would have a life worth living. The benefits such a person would enjoy do not satisfy this requirement and do therefore provide (merely) compensating reasons.

Finally, I considered some objections to this defence of the asymmetry. One of these objections was that this defence of the asymmetry is incompatible with *Q*. I argued that this is mistaken. My defence of the asymmetry is compatible with the axiological claim that it would be better if those who exist are better off. What we would have to deny on my view is rather the claim that we are required to do what would be best.

# 9. Applications and conclusion

The main focus of the thesis has been harm as a moral concept and the focus of much of the discussion has been on topics related to population ethics. However, harm is typically appealed to, not only regarding future generations, but in applied ethics, political theory and normative ethics in general. The analysis of harm which I have argued for therefore has potentially far-reaching consequences.

In this final chapter I will explore three areas where the analysis of harm and the claim that there are two kinds of reasons can be applied: the ethics of procreation, liberalism and the person-affecting view. First, however, I will restate the main claims from the previous chapters.

## 9.1 A brief summary of the thesis

In this thesis my main concern has been to defend the Harm Principle:

> *The Harm Principle*: if an act would harm someone then this is a reason against performing that act.

The defence consists of a negative and a positive part. For the negative part I argued that a common objection to the Harm Principle based on the non-identity problem does not succeed. The objection is that the Harm Principle can not explain why, for example, it is morally objectionable for a couple to have a child now rather than later if the child they could have now would suffer from some serious medical condition, while the child they could have later would be perfectly healthy. After all, the child they could have now would have a life worth living and she would not exist had the parents waited, so it cannot be worse for her that the act is not performed. We therefore have to conclude, according to this objection, that the parents would not harm their child if they do not wait and that the Harm Principle can not explain why they should wait.

A crucial premise in this argument is the Counterfactual Condition:

> *The Counterfactual Condition*: An act harms a person if and only if that person would have been better off had the act not been performed.

In chapter two I argued that the Counterfactual Condition should be rejected as an analysis of harm in a morally relevant sense because it has unacceptable consequences with respect to overdetermination. The upshot of this conclusion is that the argument against the Harm Principle, based on the non-identity problem, fails.

In chapter three I considered an alternative to the Counterfactual Condition: the Non-Comparative View. I argued that Non-Comparative Views fail to capture the plausible claim that to harm someone is to make that person's life go worse. Furthermore, Non-Comparative Views typically rely heavily on claims about the nature of well-being, something which an analysis of harm should avoid.

For the positive part I suggested an analysis of harm (chapters four to six) which I argued has plausible implications with respect to the non-identity problem. According to this Minimalist View of harm, *a* harms *b* if and only if:

> (1) *a* performs an act, $\phi$,
> (2*) *b* is in a state *S* such that *b*'s life is worse, taking *S* into account, than *b*'s life not taking *S* into account.
> (3*) if $\phi$ had not been performed then *S* would not have obtained at all, or would have obtained in a different way which is salient in the circumstances.

According to the Minimalist View it is plausible that a couple who chooses to have a child now rather than a different child later, where having a child now would result in a person coming into existence with a serious medical condition, would harm their child even if the handicap is not so severe as to make the child's life not worth living. If they do not postpone the pregnancy, then the child will be in a state which makes her life go worse in sense of (2*). Also, the couple are responsible for this state of affairs because had they postponed the pregnancy then state of affairs would not have obtained.

In chapters seven and eight I argued that with this analysis of harm there is a reason to appeal to harm in population ethics. I argued that a common intuition, strong asymmetry, can be defended by making a distinction between two kinds of reasons (requiring and compensating) and assigning a dual role to harm. The first role of harm is specified by the Harm Principle. If an act would harm someone then this is a reason not to perform the act. The second role, I suggested, is captured by the following Principle of Permissibility:

> *The Principle of Permissibility*: an act is permissible if (i) it is favoured by the balance of reasons *or* (ii) it does no harm.

According to this principle, it is always permissible to perform an act which does no harm, even if it is not favoured by the balance of reasons. I suggested that this normative claim captures the importance of autonomy. The second role of harm is then that harm sets the limits within which a person is free

to pursue her own aims even when doing so would not be favoured by the balance of reasons.

The Principle of Permissibility is also congenial with the distinction between two kinds of reasons. Because it is permissible to perform sub-optimal acts which do no harm, we can say that the benefits in favour of an optimal alternative in such cases only provide compensating reasons because these benefits are not preventions of harm. Preventing harm, on this view, functions as a precondition for benefits to provide requiring reasons. By attaching this importance to harm it is possible to defend the asymmetry as an instance of a moral option grounded in autonomy.

## 9.2 Procreative freedom

A consequence of the dual role of harm which I introduced in order to defend the asymmetry is that people have considerable procreative freedom. According to my defence of the asymmetry it cannot be required to create a person because of the benefits she would enjoy since these benefits only provide compensating reasons, not requiring ones. Assuming that there are no further relevant considerations besides harm and benefit, the only way it could be required to create a person is then if doing so would benefit other people or if it would prevent harm.

Defending procreative freedom in this way has interesting consequences regarding selection and enhancement of future people. Regarding selection, it is very likely that we will to a great extent be able to select children on the basis of whether they are more likely to have certain traits because of their genetic makeup. The moral question with respect to this practice is whether it is permissible to effectively see to it that people with certain traits are not born. The obvious worry about this practice is that it is elitistic, or "eugenic", and that we should therefore not try to prevent people from being born with certain traits.

The analysis of harm, and the limits to procreative freedom, which I have argued for in this thesis implies that there is a moral ground for preventing people from being born with certain traits. If those traits make a negative contribution to a person's life then there is a reason against creating that person. However, the view I have been arguing for does not imply that any trait which would be detrimental to a future person's well-being suffices for an obligation not to create that person. If the person would enjoy benefits which would compensate for the harmful trait then the choice to create this person would be within a couple's procreative freedom.

Regarding enhancement it is sometimes claimed that there is a morally relevant difference between enhancements and treatments. We are not morally required to provide enhancements which would make people better off but we are required to develop treatments for harmful conditions. To emphasise treat-

ments instead of enhancements can also be a way of arguing against selection: instead of selecting people who lack certain traits we should provide treatment to people with these traits.

That there is a morally relevant difference between enhancements and treatments is however hard to maintain. One difficulty lies in that the distinction between enhancement and treatment seems to presuppose that there is a morally relevant distinction to be made between normal and abnormal traits or conditions. The very notion of a "treatment" suggests that there is a disease, or abnormality, involved while the notion of "enhancement" suggests that the person to receive the enhancement passes for "normal". But, it is unclear why it should matter whether a condition, shortness for example, is the consequence of a genetic abnormality or of "natural" genetic variation.[1] To give another example, it is unclear why it would be more important to treat a depression when the cause is a abnormal, say a bipolar condition, rather than when the depression is caused by something more normal, such as stress. This places an unwarranted emphasis on irrelevant conditions, namely, whether the *cause* of one's shortness or depression is natural or not. What matters, at least from a moral point of view, is whether and how shortness or depression affects the quality of one's life and this is in general quite independent of what caused these conditions.

There are also pragmatic reasons against making the distinction part of common practice. In so far as the distinction between enhancement and treatment relies on some traits being "natural" and it therefore reinforces the view that people who suffer from abnormal conditions, those that need "treatment", are defective in some way and should be fixed.

In this context, the distinction between two kinds of benefits can perhaps serve as a replacement for the enhancement/treatment distinction. On my view, there is a difference between enhancements which are preventions of harm and those which are not. The former kind, "pure" enhancements, are like "pure" benefits in that they are not sufficient to ground a moral requirement. We are therefore not morally required to develop enhancements in cases where a person who would enjoy the enhancement would not have been harmed otherwise. However, we do have good moral reasons for developing enhancements which would prevent harm. At the moment it is of course difficult to say where this distinction is to be drawn, partly because I have remained neutral with respect to the details of what makes life worth living, or well-being. However, the distinction between two kinds of benefits can capture what is plausible in the enhancement/treatment distinction.

A typical example of procreative freedom which deserves comment is abortion. My analysis of harm might seem to imply that in many cases a foetus would be harmed by being aborted. In cases where the foetus would have developed into a human being with a life worth living, had the abortion not been

[1] See also Buchanan et al. (2001, pp. 110–15) who argue against the distinction on similar grounds. They do however defend a more limited use of the distinction on pragmatic grounds.

performed, then the foetus would have developed into a person with a life worth living then performing the abortion would be, it could be argued, worse for the foetus. It could perhaps be objected that the foetus is not the same person as the grown human, and that aborting a foetus is not to harm *that* person.[2] However, as Luper (2009, pp. 202–3) argues, the foetus is certainly the same animal as the grown human. Whether or not the foetus is the same person as the human is therefore quite irrelevant to whether abortion harms the foetus.

The difficult question regarding abortion is therefore not whether the foetus is a person or not but whether it is the kind of entity which can be harmed. The Minimalist View of harm is neutral with respect to this question. It is uncontroversial that persons can be harmed, but when we consider other kinds of entities there is considerable room for disagreement. Animals, collectives of people, social institutions or even inanimate objects are all candidates for being proper harm subjects. The Minimalist View does place some constraint on what can plausibly be harmed because it is formulated in terms of well-being. Only things which can be said to have well-being are therefore possible subjects of harm on this view. However, this still leaves it open whether all animals, for example, can be harmed in a morally relevant sense.

## 9.3 Liberalism

The Harm Principle, as I understand it, is a moral principle specifying one source of moral reasons. However, it is sometimes claimed that harm also plays an important role in political theory, especially in liberal views. J. S. Mill famously claimed that "the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others" (Mill 2003, p. 94) and many liberals have followed Mill in this respect. Of course, liberals are not the only ones who think that harm is one ground for justified coercion. What characterises the liberal view is that harm is the *only* ground for justified coercion.[3]

Liberalism, thus conceived, is typically thought to have two features. First, liberalism is *anti-paternalistic* because it is only justified to limit a person's liberty in order to prevent harm to *others*. On the liberal view, there are things which are not the state's business and what an individual chooses to do with him or herself is one of those things.[4] Secondly, liberalism is supposed to be *anti-moralistic* in the sense that the state is not justified to enforce public opinion or commonly held values for their own sake. It is not enough that

[2] McMahan (1988, p. 54), for example, claims that a foetus and the person this foetus would develop into are not the same person because there is no psychological continuity between the two.

[3] See Feinberg (1987, pp. 10–16, 26–27).) for an overview of other so-called "liberty limiting principles" and how they relate to liberalism.

[4] See Dworkin (1997).

an act would be offensive to public sentiment or morality for the state to be justified in exercising coercion.[5]

Can a liberal view which preserves these two features be formulated on the basis of the analysis of harm presented in this thesis? Consider the anti-moralistic feature first. In a sense, my analysis of harm is trivially compatible with this feature of liberalism because I have merely been arguing for an analysis of harm and have left the relevance of any non-harmful effects, such as offences, to one side. Also, I have not analysed harm in terms of public opinion or sentiments so Mill's principle, interpreted as I have suggested we should understand harm, would not justify coercion in order to enforce public opinion or commonly held values. However, the Harm Principle as a liberty limiting principle is clearly not "anti-moralistic" in the sense that the state is justified in enforcing good, or moral, behaviour. Many instances of enforced beneficence, charity taxes or so-called "bad samaritan laws" for example, would on the analysis argued for in this thesis be preventions of harm and therefore sanctioned by a liberty-limiting principle based on harm to others.

However, liberals tend to differ with respect to how much emphasis they place on the anti-moralistic part of their view. Feinberg (1987, ch. 4), for example, argues that bad samaritan laws are in principle sanctioned by the most plausible liberal view though there might be pragmatic reasons for restricting their use. What is important about the anti-moralistic part of liberalism is, presumably, that members of society should not be coerced on the basis of public opinion. That is, the liberal is not asserting that the state is not allowed to use coercion when there are good moral reasons for doing so.[6]

Turning to the second feature of liberalism, anti-paternalism, it should be clear that the Harm Principle and the Minimalist View of harm have paternalistic implications. As I formulated the Harm Principle it does not make a difference between harm to oneself and harm to others. Basing the limits of justified coercion on this version of the Harm Principle would therefore imply that the state is justified in coercing individuals for their own good.

Suppose that there are other reasons for limiting coercion to preventing harm to others. It could be argued, for example, that there are pragmatic reasons for not giving the state the authority to coerce people for the sake of preventing harm to oneself. However, the analysis of harm which I have argued for would allow paternalistic coercion but in a more indirect way since, as was noted above, my view allows bad samaritan laws. Consider for example child labour in sweatshops. Even if the state would not be allowed to prevent children from applying for work in such places in order to prevent harm to the

[5] This point is emphasised by Hart (1963).

[6] See also Hart (1963, pp. 21–24). Hart makes a distinction between public opinion and critical morality ("positive" and "critical" morality in Hart's words) and argues that coercion based on public morality is unjustified by the liberal's standards. See also Dworkin (1977, pp. 254–5) for similar remarks. For Hart and Dworkin, it seems, the important question for the liberal is not whether morality can justifiably be enforced as such but whether mere public opinion can be.

children, it would be justified in coercing the employers to not hire children as workers. This would amount to paternalistic coercion since the reason for restricting the employers freedom is to prevent harm to potential employees.

There is one form of paternalism which my view does not endorse. If we consider the distinction between requiring and compensating reasons we can say that coercing individuals to bestow benefits of the latter kind is not justified. This restriction creates a sphere of personal freedom which is not the state's business. As I have argued, procreative decisions are often within this sphere but liberals typically think that it is larger than that. It should be emphasised however that harm-prevention, on any plausible liberal view, is not a sufficient reason for legitimate state coercion but only necessary. That is, one could claim that many instances of state coercion which prevents harm is illegitimate but for reasons which are unrelated to harm. This suggests that coercion constrained by a liberty-limiting principle of Mill's kind does not establish a sphere of personal freedom, nor does it protect against paternalism, to the extent which it has been thought to do.[7]

## 9.4 The person-affecting view

In chapter one I argued that the Harm Principle belongs to a class of normative theories which are commonly classified as "person-affecting". Exactly how to formulate the person-affecting view is however an unclear and contested matter. The general, though imprecise, idea is that whether an act is wrong or morally objectionable depends of how the act would affect people. Many philosopher's have found the person-affecting view plausible. Scanlon (1998), for example, claims that the morality of right and wrong (morality in a "narrow sense") is characterised by what we owe to each other. Wrongness, on Scanlon's view, is constrained by person-affecting considerations since an act cannot be wrong unless it fails to conform to the standard of what we owe to other people – whatever that standard in the end is and whoever is to count as "other people".[8] Another example is Roberts who in a series of publications (2003*a*, 2003*b*, 2010) has developed and defended the claim that an act is wrong only if it *wrongs* someone.

The non-identity problem is sometimes claimed to be a devastating objection to such views and that we should therefore abandon the person-affecting

[7] Holtug (2002) argues for a similar conclusion.

[8] See Scanlon (1998, pp. 172–3, 177–87, 229). It should be noted however that Scanlon seems to suggest a slightly different view at times. For example, he claims that when considering whether an act is a part of what we owe to each other "we cannot envisage the reaction of every actual person. We can consider only representative cases" (p.171). Whether this is just a pragmatic device which we have to make use of because of our limited cognitive capacities or whether Scanlon thinks that this is a more principled part of morality is however unclear. Parfit (2003) argues that Scanlon's theory can be formulated without any commitment to the person-affecting view.

approach in favour of a neutral approach, for example Parfit's principle *Q*.[9] As I have argued in this thesis this is not the case: we can explain the moral objection to the young girl's choice by appealing to the harm she would be responsible for. This conclusion is significant to population ethics because it puts person-affecting views on a par with neutral views when it comes to same-number cases like the non-identity problem.

We should not overestimate the consequences of appealing to the person-affecting view for two reasons. First, I do not claim to have put all views which, at some time or other, have been called "person-affecting" on a par with theories which are neutral with respect to identity. Some *versions* of the person-affecting view gain no credibility from the defence of the Harm Principle presented in this thesis. For example, a person-affecting view according to which an act is wrong only if it makes someone worse off than they would otherwise have been is, it seems to me, refuted by the non-identity problem. Axiological versions of the person-affecting view which use the same restriction, i.e., that *A* is worse than *B* only if *A* is worse than *B* for someone, are also not supported by my defence of the Harm Principle. Regarding these version of the person-affecting view I would therefore agree with their critics and say that morality is not person-affecting *in that sense*. Second, as was also mentioned in chapter one, we can distinguish between cases where the same number of people will exist whatever we do, same-number cases, and cases where a different number of people will exist depending on what do, different-number cases. Cases of the latter kind pose new and comparatively more difficult problems for ethical theory than same-number cases, the most famous being the repugnant conclusion.[10] Since I in this thesis have mainly been concerned with same-number cases it would be premature to claim that harm, or the person-affecting view, can help us solve all of the puzzles of population ethics.

One puzzle which I have argued can be solved is the asymmetry. This solution relied on there being two kinds of reasons and in particular that some benefits only provide compensating reasons. This approach to harms and benefits does however seem to lead to the repugnant conclusion. What my defence of the asymmetry amounts to is that even though it would be better, by any reasonable axiological standard, to add a person to a population it does not follow that it is morally required to add that person. That is, a very large population where everyone lives on "muzak and potatoes" might be better than a

[9] See for example Broome (2004, p. 136). However, Broome only claims that the non-identity problem refutes axiological versions of the person-affecting view. Parfit (2011, pp. 217–43) argues that Scanlon's version of the person-affecting view should be rejected partly because of the non-identity problem.

[10] "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" (Parfit 1984, p. 388).

smaller population where people enjoy Mozart, fine cuisine and the poetry of William Blake. If the benefits to the additional people in the large population are of the compensating kind then we are not required to choose the larger population over the smaller one. However, suppose that the two populations represent possible states of the world in the far future and that they are completely disjoint: no one who exists in the larger population would exist in the smaller population. The solution I have suggested to the asymmetry therefore indicates that it would be permissible to choose either. This, some would say, is counterintuitive. It is not enough that we are allowed to opt for the smaller population, we *should* do so. The claims made herein do not determine how to go about solving this problem. All that we can say, so far, is that a person-affecting view, based on the Harm Principle, is over the first hurdle in that it is adequate for same-number cases. But, there is a lot more work to be done.

# Bibliography

Åkvist, L. (1968), 'Chisholm-Sosa logics of intrinsic betterness and value', *Noûs* **2**(3), 253–270.

Arrhenius, G. (2000), *Future Generations: A challenge for moral theory (Dissertation)*, Uppsala: Acta Universitatis Upsaliensis.

Arrhenius, G. (2003), 'The person-affecting restriction, comparativism, and the moral status of potential people', *Ethical Perspectives* **10**, 185–195.

Arrhenius, G. (2008), 'Life extension versus replacement', *Journal of Applied Philosophy* **25**(3), 211–227.

Arrhenius, G. (2009), Can the person affecting restriction solve the problems in population ethics?, *in* M. A. Roberts & D. T. Wasserman, eds, 'Harming Future Persons', Dordrecht: Springer.

Arrhenius, G. (2013), Population Ethics: the Challenge of Future Generations. Manuscript.

Arrhenius, G. & Rabinowicz, W. (2010), Better to be than not to be, *in* H. Joas & B. Klein, eds, 'The Benefit of Broad Horizons', Leiden: Brill.

Bayles, M. D. (1976), 'Harm to the unconceived', *Philosophy & Public Affairs* **5**(3), 292–304.

Belshaw, C. (2003), 'More lives, better lives', *Ethical Theory and Moral Practice* **6**, 127–141.

Benatar, D. (2006), *Better Never to Have Been: the Harm of Coming Into Existence*, Oxford: Clarendon Press.

Bigelow, J. & Pargetter, R. (1988), 'Morality, potential persons and abortion', *American Philosophical Quarterly* **25**(2), 173–181.

Bradley, B. (2004), 'When is death bad for the one who dies?', *Nous* **38**(1), 1–28.

Bradley, B. (2009), *Well-Being and Death*, Oxford: Clarendon Press.

Bradley, B. (2012), 'Doing away with harm', *Philosophy & Phenomenological Research* **85**(2), 390–412.

Bradley, B. (2013), 'Asymmetries in benefiting, harming and creating', *Journal of Ethics* **17**(1-2), 37–49.

Braham, M. & van Hees, M. (2012), 'An anatomy of moral responisbility', *Mind* **121**(483), 601–634.

Bratman, M. E. (1992), 'Shared cooperative activity', *The Philosophical Review* **101**(2), 327–341.

Brogan, A. P. (1919), 'The fundamental value universal', *The Journal of Philosophy, Psychology and Scientific Methods* **16**(4), 96–104.

Broome, J. (1999), *Ethics out of Economics*, Cambridge: Cambridge University Press.

Broome, J. (2004), *Weighing Lives*, Oxford: Oxford University Press.

Brown, C. (2011), 'Better never to have been believed: Benatar on the harm of existence', *Economics and Philosophy* **27**, 45–52.

Buchanan, A. (1984), 'The right to a decent minimum of health care', *Philosophy & Public Affairs* **13**(1), 55–78.

Buchanan, A., Brock, D. W., Daniels, N. & Wikler, D. (2001), *From Chance to Choice: Genetics and Justice*, Cambridge: Cambridge University Press.

Bykvist, K. (1998), *Changing Preferences: A Study in Preferentialism*, Uppsala: Acta Universitatis Upsaliensis.

Bykvist, K. (2006), 'Prudence for changing selves', *Utilitas* **18**(3), 264–283.

Bykvist, K. (2007*a*), 'The benefits of coming into existence', *Philosophical Studies* **135**, 335–362.

Bykvist, K. (2007*b*), 'Violations of normative invariance: some thoughts on shifty oughts', *Theoria* **73**(2), 264–283.

Carlson, E. (1995), *Consequentialism Reconsidered*, Dordrecht: Kluwer Academic Publishers.

Carlson, E. (1997), 'A note on Moore's organic unities', *The Journal of Value Inquiry* **31**(1), 55–59.

Carlson, E. (2002), 'Deliberation, foreknowledge, and morality as a guide to action', *Erkenntnis* **57**, 71–89.

Chisholm, R. (1963), 'Supererogation and offence: A conceptual scheme for ethics', *Ratio* **5**, 1–14.

Chisholm, R. M. (1968), 'The defeat of good and evil', *Proceedings and Addresses of the American Philosophical Association* **42**, 21–38.

Chisholm, R. & Sosa, E. (1966), 'On the logic of "intrinsically better"', *American Philosophical Quarterly* **3**(3), 244–249.

Daniels, N. (1981), 'What is the obligation of the medical profession in the distribution of health care?', *Social Science and Medicine* **15**(4), 129 – 133.

De George, R. T. (1980), The environment, rights, and future generations, *in* E. Partridge, ed., 'Responsibilities to Future Generations', Buffalo: Prometheus Books.

Driver, J. (1992), 'The suberogatory', *Australasian journal of philosophy* **70**(3), 286–295.

Dworkin, G. (1977), *Taking Rights Seriously*, Cambridge, Ma: Harvard University Press.

Dworkin, G. (1997), Paternalism, *in* G. Dworkin, ed., 'Mill's On Liberty: Critical essays', Lanham, Md: Rowman and Littlefield.

Eggleston, B. (2000), 'Should consequentialists make Parfit's second mistake? A refutation of Jackson', *Australasian Journal of Philosophy* **78**(1), 1–15.

Elliot, R. (1989), 'The rights of future people', *Journal of Applied Philosophy* **6**(2), 159–169.

Elstein, D. (2005), 'The asymmetry of creating and not creating life', *The Journal of Value Inquiry* **39**, 49–59.

Feinberg, J. (1980), The rights of animals and unborn generations, *in* E. Partridge, ed., 'Responsibilities to Future Generations', Buffalo: Prometheus Books.

Feinberg, J. (1986), 'Wrongful life and the counterfactual element in harming', *Social Philosophy and Policy* **4**(1), 145–178.

Feinberg, J. (1987), *Harm to Others*, Oxford: Oxford University Press.

Feit, N. (2002), 'The time of death's misfortune', *Nous* **36**(3), 359–383.

Feit, N. (2013), 'Plural harm', *Philosophy & Phenomenological Research* pp. 1–28.

Feldman, F. (1991), 'Some puzzles about the evil of death', *The Philosophical Review* **100**(2), 205–227.

Feldman, F. (1992), *Confrontations With the Reaper*, Oxford: Oxford University Press.

Feldman, F. (1997), *Utilitarianism, Hedonism, and Desert*, Cambridge: Cambridge University Press.

Foster, C., Hope, T. & McMillan, J. (2006), 'Submissions form non-identity claimants: the non-identity problem and the law', *Medical Law* **25**, 159–173.

Gert, J. (2000), 'Practical rationality, morality, and purely justificatory reasons', *American Philosophical Quarterly* **37**(3), 227–243.

Gert, J. (2003), 'Requiring and justifying: two dimensions of normative strength', *Erkenntnis* **59**(1), 5–36.

Gert, J. (2007), 'Normative strength and the balance of reasons', *Philosophical Review* **116**(4), 533–562.

Gilbert, M. (2006), 'Who's to blame? Collective moral responsibility and its implications for group members', *Midwest Studies In Philosophy* **30**(1), 94–114.

Hall, N. (2000), 'Causation and the price of transitivity', *The Journal of Philosophy* **97**(4), 198–222.

Hanser, M. (1990), 'Harming future people', *Philosophy & Public Affairs* **19**(1), 47–70.

Hanser, M. (2008), 'The metaphysics of harm', *Philosophy & Phenomenological Research* **77**(2), 421–450.

Hare, C. (2007), 'Voices from another world: must we respect the interests of people who do not, and will never, exist?', *Ethics* **117**(3), 498–523.

Harman, E. (2004), 'Can we harm and benefit in creating?', *Philosophical Perspectives* **18**(1), 89–113.

Harman, E. (2009), Harming as causing harm, *in* M. Roberts & D. T. Wasserman, eds, 'Harming Future Persons', Dordrecht: Springer.

Harman, J. D. (1981), 'Harm, consent and distress', *The Journal of Value Inquiry* **15**(4), 293–309.

Harris, J. (1990), 'The wrong of wrongful life', *Journal of Law and Society* **17**(1), 90–105.

Hart, H. L. A. (1963), *Law, Liberty, and Morality*, Oxford: Oxford University Press.

Hart, H. L. A. & Honoré, T. (1985), *Causation in the Law*, Oxford: Oxford University Press.

Heyd, D. (1982), *Supererogation*, Cambridge: Cambridge University Press.

Heyd, D. (1994), *Genethics: Moral Issues in the Creation of People*, Berkeley: University of California Press.

Hohfeld, W. N. (1966), *Fundamental Legal Conceptions*, New Haven and London: Yale University Press.

Holtug, N. (2001), 'On the value of coming into existence', *The Journal of Ethics* **5**, 361–384.

Holtug, N. (2002), 'The harm principle', *Ethical Theory and Moral Practice* **5**, 357–389.

Holtug, N. (2003), 'Good for whom?', *Theoria* **69**(1-2), 4–20.

Holtug, N. (2010), *Persons, Interests, and Justice*, Oxford: Oxford University Press.

Hurka, T. (1998), 'Two kinds of organic unity', *The Journal of Ethics* **2**(4), 299–320.

Hurley, P. (1995), 'Getting our options clear: A closer look at agent-centered options', *Philosophical Studies* **78**(2), 163–188.

Jackson, F. (1991), 'Decision-theoretic consequentialism and the nearest and dearest objection', *Ethics* **101**(3), 461–482.

Jackson, F. & Pargetter, R. (1986), 'Oughts, options, and actualism', *The Philosophical Review* **95**(2), 233–255.

Johansson, J. (2010), 'Being and betterness', *Utilitas* **22**(3), 285–302.

Kagan, S. (1989), *The Limits of Morality*, Oxford: Clarendon Press.

Kagan, S. (1998), *Normative Ethics*, Boulder: Westview Press.

Kahane, G. & Savulescu, J. (2008), 'The moral obligation to create children with the best chance of the best life', *Bioethics* **23**(5), 274–290.

Kahane, G. & Savulescu, J. (2012), 'The concept of harm and the significance of normality', *Journal of Applied Philosophy* **29**(4), 318–332.

Kavka, G. S. (1982), 'The paradox of future individuals', *Philosophy & Public Affairs* **11**(2), 93–112.

Korsgaard, C. (1983), 'Two distinctions in goodness', *The Philosophical Review* **92**(2), 169–195.

Kumar, R. (2003), 'Who can be wronged?', *Philosophy & Public Affairs* **31**(2), 99–118.

Lewis, D. (1973*a*), 'Causation', *Journal of Philosophy* **70**, 556–567.

Lewis, D. (1973*b*), *Counterfactuals*, Oxford: Blackwell.

Lewis, D. (1979), 'Counterfactual dependence and time's arrow', *Noûs* **13**(4), 455–476.

Lewis, D. (1986), *On the Plurality of Worlds*, New York: Basil Blackwell.

Lewis, D. (2000), 'Causation as influence', *Journal of Philosophy* **97**(4), 182–197.

Luper, S. (2004), 'Posthumous harm', *American Philosophical Quarterly* **41**(1), 63–72.

Luper, S. (2009), *The Philosophy of Death*, Cambridge: Cambridge University Press.

Mackie, J. L. (1955), 'Responsibility and language', *Australasian Journal of Philosophy* **33**(3), 143–159.

Mackie, J. L. (1980), *The Cement of the Universe*, Oxford: Oxford University Press.

Macklin, R. (1980), Can future generations correctly be said to have rights?, *in* E. Partridge, ed., 'Responsibilities to Future Generations', Buffalo: Prometheus Books.

McIntyre, A. (2001), 'Doing away with double effect', *Ethics* **111**(2), 219–255.

McMahan, J. (1981), 'Problems of population theory', *Ethics* **92**(1), 96–127.

McMahan, J. (1988), 'Death and the value of life', *Ethics* **99**(1), 32–61.

McMahan, J. (2001), Wrongful life: Paradoxes in the morality of causing people to exist, *in* J. Harris, ed., 'Bioethics', Oxford: Oxford University Press.

McMahan, J. (2009), Asymmetries in the morality of causing people to exist, *in* M. A. Roberts & D. T. Wasserman, eds, 'Harming Future Persons', Dordrecht: Springer.

McMahan, J. (2013), 'Causing people to exist and saving people's lives', *Journal of Ethics* **17**(1-2), 5–35.

Mill, J. S. (2003), *Utilitarianism and On Liberty*, Malden, MA: Blackwell.

Moore, G. E. (1993), *Principia Ethica (Revised edition)*, Cambridge: Cambridge University Press.

Nagel, T. (1986), *The View From Nowhere*, Oxford: Oxford University Press.

Nagel, T. (1991), *Mortal Questions*, Cambridge: Cambridge University Press.

Narveson, J. (1967), 'Utilitarianism and new generations', *Mind* **76**(301), 62–72.

Narveson, J. (1973), 'Moral problems of population', *The Monist* **57**, 62–86.

Norcross, A. (2005), 'Harming in context', *Philosophical Studies* **123**, 149–173.

Österberg, J. (1996), Value and existence: The problem of future generations, *in* S. Lindström, R. Sliwinski & J. Österberg, eds, 'Odds and Ends', Uppsala: Department of Philosophy, Uppsala University.

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

Parfit, D. (2003), 'Justifiability to each person', *Ratio* **16**(4), 368–390.

Parfit, D. (2011), *On What Matters: volume two*, Oxford: Oxford University Press.

Parsons, J. (2002), 'Axiological actualism', *Australasian Journal of Philosophy* **80**(2), 137–147.

Paul, L. A. (1998), 'Keeping track of the time: Emending the counterfactual analysis of causation', *Analysis* **58**(3), 191–198.

Paul, L. A. (2000), 'Aspect causation', *The Journal of Philosophy* **97**(4), 235–256.

Perry, S. (2003), 'Harm, history, and counterfactuals', *San Diego Law Review* **40**(4), 1283–1313.

Persson, I. (1997), 'Ambiguities in Feldman's desert-adjusted values', *Utilitas* **9**(3), 319–327.

Petersson, B. (2004), 'The second mistake in moral mathematics is not about the worth of mere participation', *Utilitas* **16**(3), 288–315.

Pitcher, G. (1984), 'The misfortunes of the dead', *American Philosophical Quarterly* **21**(2), 183–188.

Rabinowicz, W. (2009), 'Broome and the intuition of neutrality', *Philosophical Issues* **19**(1), 389–411.

Rabinowicz, W. & Rønnow-Rasmussen, T. (2000), 'A distinction in value: intrinsic and for its own sake', *Proceedings of the Aristotelian Society* **100**(1), 33–51.

Rabinowicz, W. & Österberg, J. (1996), 'Value based on preferences', *Economics and Philosophy* **12**, 1–27.

Rachels, S. (1998), 'Is it good to make happy people?', *Bioethics* **12**(2), 93–110.

Raz, J. (1984), 'On the nature of rights', *Mind* **93**(370), 194–214.

Raz, J. (1988), *The Morality of Freedom*, Oxford: Clarendon Press.

Reiman, J. (2007), 'Being fair to future people: the non-identity problem in the original position', *Philosophy & Public Affairs* **35**(1), 69–92.

Roberts, M. (1998), *Child Versus Childmaker*, Lanham, Md.: Rowman and Littlefield.

Roberts, M. (2003*a*), 'Can it ever be better never to have existed at all? Person-based consequentialism and a new repugnant conclusion', *Journal of Applied Philosophy* **20**(2), 159–185.

Roberts, M. (2003*b*), 'Is the person-affecting intuition paradoxical?', *Theory and Decision* **55**, 1–44.

Roberts, M. (2010), *Abortion and the Moral Significance of Merely Possible Persons*, Dordrecht: Springer.

Roberts, M. (2011), 'The asymmetry: a solution', *Theoria* **77**, 333–367.

Roemer, J. E. (1993), 'A pragmatic theory of responsibility for the egalitarian planner', *Philosophy & Public Affairs* **22**(2), 146–166.

Ross, W. D. (2002), *The Right and the Good*, Oxford: Clarendon Press.

Sartorio, C. (2004), 'How to be responsible for something without causing it', *Philosophical Perspectives* **18**, 315–336.

Savulescu, J. (2001), 'Procreative beneficence: why we should select the best children', *Bioethics* **15**(5/6), 413–426.

Scanlon, T. M. (1998), *What we Owe to Each Other*, Cambridge M.A: Harvard University Press.

Scanlon, T. M. (2008), *Moral Dimensions: Permissibility, Meaning, Blame*, Cambridge M.A: Harvard University Press.

Schaffer, J. (2003), 'Overdetermining causes', *Philosophical Studies* **114**, 23–45.

Scheffler, S. (2003), *The Rejection of Consequentialism (Revised edition)*, Oxford: Clarendon Press.

Shiffrin, S. (1999), 'Wrongful life, procreative responsibility, and the significance of harm', *Legal Theory* **5**, 117–148.

Sidgwick, H. (1981), *The Methods of Ethics*, Cambridge: Hackett Pub Co Inc.

Singer, P. (1993), *Practical Ethics*, Cambridge: Cambridge University Press.

Smilansky, S. (2007), *10 Moral Paradoxes*, Oxford: Blackwell.

Smith, M. (1994), *The Moral Problem*, Oxford: Blackwell.

Steinbock, B. (1986), 'The logical case for "wrongful life"', *The Hastings Center Report* **16**(2), 15–20.

Steinbock, B. (2009), Wrongful life and procreative decisions, *in* M. A. Roberts & D. T. Wasserman, eds, 'Harming Future Persons', Dordrecht: Springer.

Stroud, S. (1998), 'Moral overridingness and moral theory', *Pacific Philosophical Quarterly* **79**, 170–189.

Sumner, W. (1996), *Welfare, Happiness, and Ethics*, Oxford: Clarendon Press.

Temkin, L. S. (1987), 'Intransitivity and the mere addition paradox', *Philosophy & Public Affairs* **16**(2), 138–187.

Temkin, L. S. (1993), *Inequality*, Oxford: Oxford University Press.

Thomson, J. J. (2011), 'More on the metaphysics of harm', *Philosophy & Phenomenological Research* **82**(2), 436–458.

Tännsjö, T. (1989), 'The morality of collective actions', *The Philosophical Quarterly* **39**(155), 221–228.

Urmson, J. O. (1958), Saints and heroes, *in* A. I. Melden, ed., 'Essays in Moral Philosophy', Seattle: University of Washington Press.

Vanderheiden, S. (2006), 'Conservation, foresight, and the future generations problem', *Inquiry* **49**(4), 337–352.

Waldron, J. (1984), Introduction, *in* J. Waldron, ed., 'Theories of Rights', Oxford: Oxford University Press.

Williams, B. (1984), A critique of utilitarianism, *in* J. J. C. Smart & B. Williams, eds, 'Utilitarianism: For and Against', Cambridge University Press.

Woodward, J. (1986), 'The non-identity problem', *Ethics* **96**(4), 804–831.

# Index