



# LUND UNIVERSITY

## Reasons, Blame, and Collective Harms

Gunnemyr, Mattias

2021

*Document Version:*

Publisher's PDF, also known as Version of record

[Link to publication](#)

*Citation for published version (APA):*

Gunnemyr, M. (2021). *Reasons, Blame, and Collective Harms*. [Doctoral Thesis (compilation), Department of Philosophy, Joint Faculties of Humanities and Theology]. Lund University (Media-Tryck).

*Total number of authors:*

1

### General rights

Unless other specific re-use rights are stated the following general rights apply:

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: <https://creativecommons.org/licenses/>

### Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117  
221 00 Lund  
+46 46-222 00 00

# Reasons, Blame, and Collective Harms

MATTIAS GUNNEMYR

DEPARTMENT OF PHILOSOPHY | LUND UNIVERSITY





## Reasons, Blame, and Collective Harms



# Reasons, Blame, and Collective Harms

Mattias Gunnemyr



**LUND**  
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Joint Faculties of Humanities and Theology,  
Lund University, Sweden.

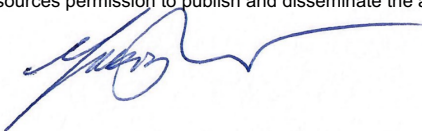
To be defended at LUX C:121, Saturday 2021-11-20 at 10 a.m.

*Faculty opponent*  
Stephanie Collins (ACU Melbourne)

<b>Organization</b> LUND UNIVERSITY		<b>Document name:</b> DOCTORAL DISSERTATION
Author: Mattias Gunnemyr		<b>Date of issue:</b> 2021-11-20
		<b>Sponsoring organization</b>
<b>Title and subtitle:</b> Reasons, Blame, and Collective Harms		
<p><b>Abstract:</b> Collective harm cases are situations in which things will become worse if enough acts of a certain kind are performed but no single act of the relevant kind will make a difference to the outcome. The inefficacy argument says that since one such act does not make a difference to the outcome, you have no outcome-related reason to refrain from acting in this way. If this argument holds, you have no climate-change-related reason to refrain from going for a drive in a fossil fuel powered car, and no harm-to-the-victim-related reason to refrain from flipping the switch in Derek Parfit's (1984) famous case of the harmless torturers. There are two ways in which you could understand the inefficacy argument. Either, it says that you lack a reason to act in the relevant way because one such act makes no difference at all to the outcome, or it says that you lack a reason to act in the relevant way because the outcome will occur whether or not you act in this way. Either way, the argument is unfounded. Acting in the relevant way <i>does</i> make a difference to the outcome. Given that there is a possibility that the outcome will occur and a possibility that it will not, acting in the relevant way makes the outcome closer to happening (or further from not happening). In technical terms: acting in the relevant way makes the outcome more secure within the relevant possibility horizon. Thus, the first suggested interpretation of the inefficacy argument is unsound.</p> <p>The second interpretation rests implicitly on a flawed understanding of causation according to which causes always make a difference to whether or not their outcomes occur. An improved account of causation entails that there is a causal connection between the single act and the outcome in collective harm cases. It entails, for instance, that going for just one drive in a fossil fuel powered car is a cause (one of many) of climate change, and that flipping a switch is a cause (one of many) of the victim's pain in the case of the harmless torturers. Drawing from this account of causation, it is possible to explain when, and why, you have outcome-related reasons in collective harm cases. You have an outcome-related reason to act in a certain way when acting in this way makes a good outcome more secure within the relevant possibility horizon. This account captures the intuitive idea that you have outcome-related reasons to contribute to good outcomes, and to refrain from contributing to bad ones. It also produces intuitively correct verdicts about what outcome-related reasons you have in many different kinds of cases, including collective harm cases (with or without a threshold), pre-emption cases, switching cases, overdetermination cases, omission cases, Frankfurt-style cases, cases where we disregard irrelevant possibilities, the difficult case of the thirsty traveller, and more. Importantly, this account provides the resources to pinpoint the problem in the second variety of the inefficacy argument. You might have an outcome-related reason to refrain from acting in the relevant way in collective harm cases even if the harmful outcome will occur whether you refrain or not: you have such a reason if there is a possibility that the outcome will occur, a possibility that it will not occur, and acting in this way makes the outcome closer to happening.</p> <p>There is also a question of whether you could be blameworthy for the outcome in collective harm cases. An adjusted version of the inefficacy argument says that you cannot be blameworthy for the outcome in collective harm cases since what you do makes no difference to the outcome. Also this version of the argument is mistaken, and for the same reasons. Building on the mentioned account of causation, it is possible to explain when and why you are blameworthy for an action, omission or outcome. You are blameworthy for X – where X is an act, omission or outcome – if and only if a poor quality of will of yours in relation to X was a cause of X. Like the proposed account of reasons, the account of blameworthiness produces intuitively correct verdicts in a wide range of cases.</p>		
<b>Key words:</b> Reasons, blameworthiness, causation, the inefficacy problem, helping, causal contributions, imperceptible harm, non-threshold cases, moral luck, process-connection, security-dependence, contrastive reasons, the thirsty traveller		
Classification system and/or index terms (if any)		
Supplementary bibliographical information		<b>Language:</b> English
<b>ISSN and key title</b>		<b>ISBN:</b> 978-91-89213-95-1 (print) 978-91-89213-96-8 (digital)
Recipient's notes		<b>Number of pages:</b> 338 Price
		Security classification

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature



Date 2021-09-31

# Reasons, Blame, and Collective Harms

Mattias Gunnemyr



**LUND**  
UNIVERSITY



Cover photo by Mattias Gunnemyr

Copyright pp. 1-111, 141-161, 187-211 and 245-338: Mattias Gunnemyr.

Paper 1 © Caroline Touborg and Mattias Gunnemyr (Unpublished manuscript).

Paper 2 © Inquiry, Taylor & Francis. Open access.

Paper 3 © Caroline Touborg and Mattias Gunnemyr (Unpublished manuscript).

The Joint Faculties of Humanities and Theology  
Department of Philosophy

ISBN 978-91-89213-95-1 (print)

ISBN 978-91-89213-96-8 (digital)

Printed in Sweden by Media-Tryck, Lund University  
Lund 2021



Media-Tryck is a Nordic Swan Ecolabel  
certified provider of printed material.  
Read more about our environmental  
work at [www.mediatryck.lu.se](http://www.mediatryck.lu.se)

**MADE IN SWEDEN** 

*To Janna, for love and support.  
To Arvid and Vera, for being you.  
To Lena, for always believing in me.  
To Hasse, for teaching me never to take a shortcut across a lawn.*

# Acknowledgements

According to the stereotype, philosophers work in solitude and isolation. Even though I have been working in isolation during the last year and a half due to the pandemic, I would not say that the stereotype is accurate. Without my colleagues, family and friends, this book would not be what it is, and writing it would not have been nearly as rewarding as it turned out to be. My first heartfelt thanks go out to the people who make sure that the ordinary day is never dull and never predictable. To all my colleagues at the Department of Philosophy, thank you!

Next, to my supervisors, Björn Petersson and Dan Egonsson. You helped me gather my thoughts after presenting at the Higher Seminar as a freshly baked PhD student, you let me try out different directions of the project, and you let me abandon them when they did not work out. Even more importantly, you always made me feel welcome. Dan, the first thing you did when I began my PhD studies was to invite me to dinner together with Cathrine Felix and Ben Bramble at your wonderful apartment at Stålbrogatan 2. We have had several dinners since then, at Stålbrogatan, at your cottage in Blekinge, and at my place. I appreciate all of them. Björn, we have been to conferences and workshops around half the world (almost): The Hague, Boston, Tampere, Pavia and Vienna. We also arranged *The Fifth Conference of the European Network on Social Ontology* in Lund, together with the *Metaphysics and Collectivity Group*. These have been great times! Moreover, the visits to your, Ulrika's and Vide's place at Ven have become a summer highlight, not just for me, but for my whole family. I am already looking forward to next summer's visit. Last but not least, Dan, I have always appreciated our discussions, formal and informal. They have been of great benefit to me. Björn, thanks for relentlessly reading and giving thoughtful and constructive input on all the texts I have sent to you during these years. I have lost track over how many times you have read and commented on some of the chapters in the thesis. It would not be what it is without you. *Sine qua non*.

To Gunnar Björnsson, my opponent at the final seminar. You gave me invaluable comments on what I thought was the penultimate draft of the dissertation. (It turned out that it was not.) Your comments before, during and after the seminar have helped me improve the dissertation immensely. Not to mention all the philosophical discussions we have had during conferences and workshops and in email conversations. Gunnar, I am extremely grateful to you. The same thing must be said about Caroline Touborg. Her thorough readings of most of my chapters have been central in developing this thesis. And even more importantly, Caroline, our writing sessions are some of the best philosophical moments I have had.

My thanks go to quite a few people, for quite a few things. I have people to thank for each and every chapter of this thesis. I will start with Chapter 1. I have presented drafts of this chapter at the Doctoral Seminar in Philosophy, Lund, and I wish to

thank the participants at the seminar. In particular, I wish to thank Anton Emilsson, Max Minden Ribero, Jakob Werkmäster, and Marta Johansson Werkmäster for constructive comments. I also wish to thank Gunnar Björnsson, Dan Egonsson, Janna Lundberg and Björn Petersson for important input.

Next, I have presented drafts of Chapter 2 at the Higher Seminar in Practical Philosophy here in Lund, and at the conference *Social Ontology 2021* in San Diego (online). I wish to thank the participants for valuable comments, in particular Henrik Andersson, Gunnar Björnsson, Eric Carlson, Dan Egonsson, Magnus Jedenheim Edling, Jens Johansson, Samuel Lee, Wlodek Rabinowicz, Toni Rønnow-Rasmussen, Alex Velichkov and Jakob Werkmäster. I am also grateful to Dan Egonsson, Janna Lundberg, Björn Petersson and Jakob Werkmäster for their close reading and valuable comments on earlier drafts of the chapter, and to Frits Gåvertsson for our discussions which eventually inspired me to write the section on virtue ethics. I hope you do not disagree too much with the result.

I wish to thank the participants of an informal LGRP<sup>1</sup> workshop held at the island of Ven in February 2019 for valuable discussions on a text that eventually resulted in Chapter 3. It was a fantastic workshop! Thanks Gunnar Björnsson, Olle Blomberg, Björn Petersson, András Szigeti, Wlodek Rabinowicz, Paul Russell, Matt Talbert, Caroline Touborg and Marta Johansson Werkmäster. I have also presented a draft of this chapter at the Doctoral Seminar in Philosophy, Lund, where I got many valuable comments. My thanks go to, in particular, Anton Emilsson, Jiwon Kim, Robert Pål-Wallin, Caroline Touborg, Alex Velichkov and Marta Johansson Werkmäster. In addition, I wish to thank Björn Petersson for reading and commenting on several different drafts of the chapter. I appreciate your comments.

Caroline Touborg and I elaborated many of the thoughts of Chapter 4 together. Parts of what is now Chapter 4 used to be parts of what later became “Reasons for Action” (Chapter 5). Thanks, Caroline! I am also grateful to Marta Johansson Werkmäster and Jakob Werkmäster for their comments on an early version of the chapter.

Chapter 5 “Reasons for Action” consist in a paper that Caroline Touborg and I have written together. We are grateful to audiences at the Higher Seminar in Practical Philosophy at Lund University, and at the workshop *Collective and Shared Responsibility* at the MANCEPT workshops in Manchester 2019. We would especially like to thank David Alm, Henrik Andersson, Olle Blomberg, Stephanie Collins, Dan Egonsson, Anton Emilsson, Frank Hindriks, Jakob Werkmäster, Niels de Haan, Ingvar Johansson, Björn Petersson, Wlodek Rabinowicz, Toni Rønnow-Rasmussen, David Shoemaker, Matt Talbert, Marta Johansson Werkmäster, and Bill Wringe. We would also like to thank Gunnar Björnsson and Samuel Lee for their

---

<sup>1</sup> LGRP stands for “the Lund Gothenburg Responsibility Project”.

close reading and valuable comments, which led us to revise many aspects of the paper.

Earlier versions of “Making a Vague Difference” (Chapter 8) have been presented at the Higher Seminar in Practical Philosophy at the Department of Philosophy, Lund University (2018), at the Doctoral Seminar at the same department (2020), and at the workshop *Group Agency and Collective Responsibility* in Vienna (2019). I wish to thank the participants at these events for helpful comments, and others who have contributed to the paper. In particular, I wish to thank Franz Altner, Henrik Andersson, Gunnar Björnsson, Olle Blomberg, Stephanie Collins, Dan Egonsson, Anton Emilsson, Andrés Garcia, Carol Gould, Niels de Haan, Marianna Leventi, Ingvar Johansson, Martin Jönsson, Grace Paterson, Herlinde Pauer-Studer, Björn Petersson, Matthew Rachar, an anonymous referee, Max Minden Ribeiro, Hans Bernhard Schmid, András Szigeti, Caroline Touborg, Alexander Velichkov, Jakob Werkmäster and Marta Johansson Werkmäster. This chapter is forthcoming as a standalone paper in *Inquiry*.

I want to thank Olle Blomberg for his close reading and valuable comments on Chapters 7 and 9 at a late stage in the writing process. Moreover, Björn Petersson, Matt Talbert and Caroline Touborg have all given great input on Chapter 10 that helped me clarify the argument I make in this chapter.

“You Just Didn’t Care Enough” (Chapter 11) is a joint work with Caroline Touborg. We are grateful to audiences at the *Group Agency and Collective Responsibility Workshop* in Flensburg 2019, at the *Group Agency and Collective Responsibility Workshop* in Gothenburg 2020, at the Higher Seminar in Practical Philosophy, Lund University 2020, and at the conference *Social Ontology 2020*. We would especially like to thank Franz Altner, Gunnar Björnsson, Olle Blomberg, Dan Egonsson, Anton Emilsson, Frits Gåvertsson, Niels de Haan, Frank Hindriks, Giulia Lasagni, Samuel Lee, Marianna Leventi, Carlos Nunez, Grace Paterson, Andrew Peet, Björn Petersson, Matthew Rachar, Wlodek Rabinowicz, Toni Rønnow-Rasmussen, David Schweikard, Matthew Talbert, Alexander Velichkov, Marta Johansson Werkmäster, Jakob Werkmäster and Michael Wilby. We would also like to thank Gunnar Björnsson, Anton Emilsson, and Per-Erik Milam for their close readings and valuable comments.

I am grateful to Henrik Andersson, Björn Petersson and Caroline Touborg for helpful comments on earlier versions of Chapters 12 and 13, and I wish to thank Gunnar Björnsson, Björn Petersson and Caroline Touborg for constructive comments on Chapter 14, and Dan Egonsson for more destructive ones. Both kinds of comments were much needed. I am also grateful to the participants of the *Causation and Responsibility* workshop in Bern 2021 (online), where I presented the second part of this chapter. In particular, I wish to thank Sarah Bernstein for input.

Special thanks go to Paul Robinson, who has proofread most of the thesis. Your keen eye and helpful suggestions have greatly improved the readability and elegance of this work. Special thanks also go to Marianna Leventi for reading the thesis just days before it went to printing, helping me to make some final corrections. I also wish to thank Gustav Hersmann for guiding me in taking and choosing the photo for the frontpage, and for editing this photo. Thanks Gustav!

Being a PhD is so much more than writing a thesis. To my wonderful colleagues David Alm, Andrey Anikin, Annah Smedberg-Eivers, Anna Cagnan Enhörning, Dan Egonsson, Fredrik Eriksson, Cathrine Felix, Jana Holsanova, Kristin Ingvarsdottir, Martin Jönsson, Jens Nirme, Tomas Persson, Björn Petersson, Wlodek Rabinowicz, Max Minden Ribero, Maximilian Roszko, Paul Russell, Toni Rønnov-Rasmussen, Eva Sjöstrand, Robin Stenwall, Andreas Stephens, András Szigeti, Matthew Talbert, Trond Arild Tjöstheim, Betty Tärning, Tobias Hansson Wahlberg, Annika Wallin, Anna Östberg. For coffee breaks, lunch breaks, small talk by the copy machine, and so much more.

To my fellow PhDs and post docs, Olle Blomberg, Ben Bramble, Eric Brandstedt, Anton Emilsson, Seyyed Mohsen Eslami, Andrés Garcia, Frits Gåvertsson, Fritz-Anton Fritzson, Jiwon Kim, Marianna Leventi, Gloria Mähringer, Signe Savén, Jeroen Smid, Melina Tsapos (I am happy that I can list you among the PhDs now Melina!), Alexander Velichkov, Robert Pál-Wallin. For discussions, advice and laughter. To Henrik Andersson, Marta Johansson Werkmäster and Jakob Werkmäster, for all hours of chess, coup and playing ping pong. For music Fridays. For coffee breaks. For gin & tonic tea. For making every day at the department special.

Last but not least, to my clever little girl Vera, to my thoughtful and kind soon-to-be-teenager Arvid, and to the love of my life Janna. For all weekday mornings and Friday evenings. For love and laughter. For keeping me sane. For everything. Words are not enough.

# Table of Contents

Part One: Reasons and Causation .....	15
<b>1. Collective Harm Cases and the Inefficacy Problem.....</b>	<b>17</b>
Accepting the Conclusion.....	20
Denying the Implication .....	21
Denying the Description.....	22
About This Thesis.....	26
<b>2. Non-Causal Responses .....</b>	<b>31</b>
Fairness.....	32
Virtue Ethics.....	39
Kantianism.....	47
Complicity .....	53
Reasons to Take Collective Action .....	60
Membership in a Group.....	62
Indirect Consequences.....	71
Conclusion.....	72
<b>3. Causal Responses.....</b>	<b>75</b>
Lewis' Simple Account .....	80
Wright's NESS Account.....	87
Conclusion.....	93
<b>4. Non-Superfluous Causes .....</b>	<b>97</b>
Nefsky on Helping.....	97
Paying Attention to Causation.....	99
Counterexamples to HELPING.....	107
Conclusion.....	110

<b>5.</b>	<b>Reasons for Action.....</b>	<b>113</b>
	Introduction .....	114
	Starting Assumptions.....	115
	Desiderata .....	117
	Finding a Middle Way.....	121
	Testing the Account.....	126
	Drops of Water .....	131
	Conclusion.....	136
	Appendix .....	137
<b>6.</b>	<b>Using REASON .....</b>	<b>141</b>
	Switching Cases.....	141
	Early and Late Pre-emption Cases .....	143
	Climate Change .....	145
	Double Prevention Cases.....	148
	Cases of Transitivity Failure .....	149
	Superfluous Contributions to the Underlying Dimension .....	151
	Wasteful Contributions.....	153
	Conclusion.....	154
<b>7.</b>	<b>Denying the Description I .....</b>	<b>155</b>
	Appealing to Empirical Evidence.....	158
<b>8.</b>	<b>Making a Vague Difference .....</b>	<b>163</b>
	Part I: The Kagan–Nefsky Debate.....	165
	Part II: Theories of Vagueness and Kagan’s Argument.....	172
	Part III: Two Other Versions of Kagan’s Argument.....	179
	Conclusion .....	184
<b>9.</b>	<b>Denying the Description II.....</b>	<b>187</b>
	No Free Lunch.....	187
	Problems in Threshold Cases .....	191
	Conclusion.....	194
	<b>Part Two: Blameworthiness and Causation .....</b>	<b>197</b>
<b>10.</b>	<b>Blameworthiness For.....</b>	<b>199</b>
	Does Scope Count for Nothing?.....	203
	Conclusion.....	211



<b>11. You Just Didn't Care Enough .....</b>	<b>213</b>
Introduction .....	214
Developing the Basic Idea .....	215
Characterising the Right Causal-explanatory Relation .....	219
Completing the Account .....	231
Blameworthiness and Frankfurt-style Cases .....	233
Blameworthiness in Collective Harm Cases .....	237
Conclusion .....	242
<b>12. Elaborating BLAMEWORTHINESS FOR .....</b>	<b>245</b>
Deviant Causation .....	246
Understanding Process-Connections .....	252
Process-Connections and NESS .....	254
Causal Contrasts .....	259
The in-Virtue-of Relation .....	262
Conclusion .....	267
<b>13. Applying BLAMEWORTHINESS FOR .....</b>	<b>269</b>
Non-threshold Cases .....	269
Climate Change .....	273
Pinned-In Sharks .....	275
A Potential Counterexample .....	284
Conclusion .....	291
<b>14. Moral Entails Causal .....</b>	<b>293</b>
Two Buttons .....	293
The Thirsty Traveller .....	303
Conclusion .....	320
<b>Reasons, Blame, and Collective Harms .....</b>	<b>321</b>
Questions for Future Research .....	326
<b>References .....</b>	<b>327</b>

# Part One

## Reasons and Causation

Collective harm cases are situations in which things will become worse if enough acts of a certain kind are performed but no single act of the relevant kind will make a difference to the outcome. The inefficacy argument says that since one such act does not make a difference to the outcome, I have no outcome-related reason to refrain from acting in this way. However, this argument is mistaken. It rests implicitly on a flawed understanding of causation according to which causes always make a difference to whether or not their outcomes occur. An improved account of causation entails that there is a causal connection between the single act and the outcome in collective harm cases. It entails, for instance, that going for just one drive in a fossil fuel powered car is a cause (one of many) of climate change. Building on this account of causation, it is possible to explain when, and why, you have outcome-related reasons in collective harm cases. You have an outcome-related reason to act in a certain way when acting in this way contributes to a good outcome; or more precisely, when it makes the good outcome more secure within the relevant possibility horizon. This account of reasons produces intuitively correct verdicts about what reasons we have to act in wide range of cases, including collective harm cases (with or without a threshold), pre-emption cases, switching cases, over-determination cases, omission cases, Frankfurt-style cases, cases where we disregard irrelevant possibilities, and more.



# 1. Collective Harm Cases and the Inefficacy Problem

*Collective harm cases* are cases where bad consequences follow if enough people act in a certain way but no single act of the relevant kind makes a difference to this outcome. Global warming might be such a case. When enough people drive cars running on fossil fuel, this leads to climate change. Still, it seems that no single drive worsens climate change. No extra floods, droughts or storms will occur as a result of my going for a drive. Similarly, overfishing might be such a case. When enough people fish, this leads to overfishing. Still, no one fish taken from the ocean seems to make a difference to the ability of fish to successfully reproduce, replenishing stocks, and so no one fish taken from the ocean could make a difference to those who depend on fishing for their livelihood.

Cases of this kind abound. During a hot summer, there might be a severe water shortage if enough people ignore advice to save water, but the shortage will be just as severe if I take a nice long shower. When enough people take their cars to work, there will be traffic jams, but the traffic jam will be there whether I take the car to work or not. If enough people buy factory-farmed chicken, scores of chickens will be hatched, raised and slaughtered under current factory-farmed conditions, but no single purchase of a chicken seems to affect how many chickens will meet this dreadful fate. When enough people cross a beautiful lawn, the lawn will lose its beauty, but no single crossing makes a difference to the way the lawn looks. And so on.

Collective harm cases pose a special problem. You might reason along the following lines when considering whether to buy a factory-farmed chicken at the supermarket: no chicken will suffer just because I buy this one chicken, so I might as well buy it. Or, you might think you have no climate-change-related reason not to go for a leisure drive with a fossil fuel powered car – to refrain from “joy-guzzling”.<sup>1</sup> Although there will be bad consequences if enough people drive, you might think, climate change will not become worse just because I joy-guzzle on any particular occasion. In general, each agent contemplating the possibility of a collective harm can argue: “things will be just as bad whether or not I act in this way, so there’s no

---

<sup>1</sup> This handy term was introduced by Kingston and Sinnott-Armstrong (2018).

point in doing otherwise”.<sup>2</sup> I will follow Julia Nefsky (2019) in calling this argument *the inefficacy argument*.

Besides collective harm cases, there are *collective benefit cases*. Take giving money to charity. When enough people make donations to an effective aid agency, people’s suffering will be alleviated, but no one donation seems to make a difference to the suffering of those people. Voting might be another example of this kind. When enough people vote for the right party, there will be good consequences, but it seems that no single vote will make any difference. The only difference between these two types of case is that, in the first, the outcome is harmful, and in the second, it is beneficial. Following Nefsky (2017), we might group them together and call them *collective impact cases*. The inefficacy argument applies to all collective impact cases.<sup>3</sup>

The inefficacy argument concerns *outcome-related reasons*. It says that you do not have climate-change-related reasons to refrain from joy-guzzling, that you lack future-suffering-of-chickens-related reasons to refrain from buying a chicken at the supermarket, and so on. Even if the inefficacy argument is sound, you might have reasons to refrain from acting in the relevant way. You could have a reason not to joy-guzzle because you have promised someone you will help them to move house, or because you have to work in order to meet an upcoming deadline. Or maybe there is a friend you want to visit rather than going for a leisure drive. What the inefficacy argument shows, if it is successful, is that you have no climate-change-related reason to refrain from joy-guzzling. More generally, it shows that you have no collective-impact-related reason to act in a certain way.

Reasons are considerations that speak in favour of some action, but they are not necessarily conclusive. Thus, you may have a climate-change-related reason to refrain from taking a fossil fuel powered car to get somewhere but also have a stronger reason to take it. It might, for instance, be the only way to get an injured person to the hospital in time for treatment.

In its most general form, the inefficacy argument says that, since your act does not make a difference to outcome O, you have no O-related reason to act in this way. One might then ask what it means for an act to make a difference to an outcome. An act can make a difference to an outcome in more than one way. It can affect whether the outcome occurs, when it occurs, or the manner in which it occurs. It may also

---

<sup>2</sup> Nefsky (2019: 2).

<sup>3</sup> I call these cases “collective harm cases” and “collective impact cases” since these names are generally recognised in the literature (see Nefsky 2012, 2015, 2017, 2019). Kutz (2000) uses the similar-sounding “unstructured collective harm” to refer to collective harm cases where the harmful outcome is not brought about by a joint action, and Kagan (2011) calls collective impact cases “collective action problems”. Collective impact cases (or some subset thereof) also go under the labels “the problem of many hands” (van de Poel 2011), “each-we dilemmas” (Parfit 2011) and “the I-We problem” (Kutz 2000).

make a difference to the causal history of an outcome, or to the way in which the outcome relates to other occurring events. To get a firm grip on what it is for an act to make a difference to an outcome, I will take this to mean that the outcome would not have occurred if the act had not been performed (at least, unless I state otherwise). This way of understanding what it is for an act to make a difference to an outcome is in line with standard consequentialist thinking. According to the standard form of consequentialism, it is irrelevant, for instance, whether what you do is part of the causal history of an outcome. What matters is whether your act makes things better or worse; and if your act makes things better or worse, it makes some benefit or harm occur that otherwise would not have occurred. Moreover, this idea of understanding what it is for an act to make a difference to an outcome accords with a standard view of causation, according to which an event causes an outcome if and only if the outcome would not have occurred in the absence of the event.<sup>4</sup>

With these clarifications, we can now formulate the inefficacy argument a little more specifically:

#### THE INEFFICACY ARGUMENT

- (i) If outcome O will occur whether you  $\varphi$  or not, you have no O-related reason to  $\varphi$ .
  - (ii) Outcome O will occur whether you  $\varphi$  or not.
- $\therefore$  You have no O-related reason to  $\varphi$ .

When considering a specific collective impact case, you might think that the inefficacy argument must be mistaken. You might think that you *do* have climate-change-related reasons to refrain from joy-guzzling, or an alleviation-of-suffering-related reason to give money to charity. You may not have this intuition in all collective impact cases but only in some. If you have the intuition, you have a *reasons intuition*. If this intuition is correct, the inefficacy argument must be unsound. The problem is to show exactly where it goes wrong. I will refer to this problem as the *inefficacy problem*.

Explaining where and when the inefficacy argument goes wrong is the task of Part One of this thesis. The aim is not to vindicate the notion that you have outcome-related reasons to act in a certain way in each and every collective harm case, but rather to examine the conditions under which you have reasons of this kind. Still, to anticipate, it turns out that you often do have such reasons.

---

<sup>4</sup> This understanding of causation is commonly called the “but-for” analysis in legal philosophy, but is also known as a simple counterfactual analysis of causation.

## Accepting the Conclusion

Some writers accept the inefficacy argument. Focusing on the case of climate change, Walter Sinnott-Armstrong (2005) argues that no individual has a climate-change-related reason to refrain from joy-guzzling, and that it is instead the government's job to do something about climate change, concluding that "It is better to enjoy your Sunday drive while working to change the law so as to make it illegal for you to enjoy your Sunday driving" (312). Iris Marion Young (2011) somewhat similarly argues that you are not blameworthy for buying clothes made under slave-like conditions since, typically, "it is not possible to identify how the actions of one particular individual, or even one particular collective agent, such as a firm, has directly produced harm to other specific individuals" (Young 2011: 96).<sup>5</sup> Instead, she suggests, you might have forward-looking responsibilities (which are strong reasons of a sort) to work with others to ameliorate the structural processes that lead to injustice within the global garment industry.

I think Sinnott-Armstrong and Young underestimate the full import of the inefficacy argument. If the argument is cogent, it turns out that very few of us have reasons to work for more progressive climate policies or for redressing unjust structural process. For instance, Sweden's climate policies would most likely be the same whether I personally work to change them or not. And, even if I could make some small improvement to Sweden's policies, climate change and its related harms would occur just the same. So, if the inefficacy argument is correct, I have no climate-change-related reason to work for more progressive climate policies in Sweden. Similarly, the working conditions in the factories where our clothes are made will be the same whether or not I personally join the international movement for just working conditions within the global garment industry. So, if the inefficacy argument is correct, I do not have a slave-like-working-conditions-related reason to join this movement.<sup>6</sup> So, to the extent Sinnott-Armstrong wants to hold on to the idea that I have a climate-change-related reason to work to make it illegal to go joy-guzzling, he should hesitate to accept the inefficacy argument. And, if you agree with Young that I have a slave-like-working-conditions-related reason to join the movement that challenges contemporary working conditions in the garment industry, you, also, should be concerned about the inefficacy argument.

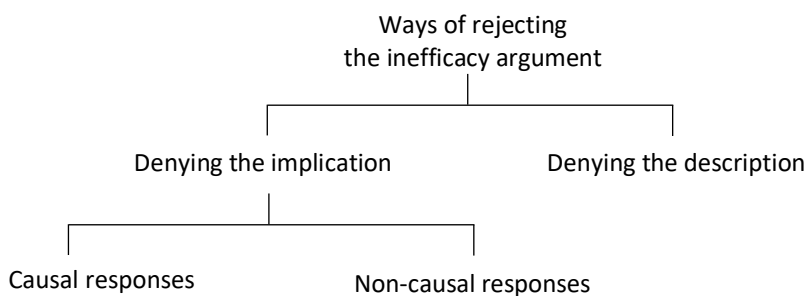
---

<sup>5</sup> This is not exactly the inefficacy argument, but it is close to it.

<sup>6</sup> The same point can be made about tracing. In most cases, it is impossible to identify how any particular action of mine produces any relevant change in the structural processes that reproduce injustice.

## Denying the Implication

Most writers reject the inefficacy argument. There are two fundamental ways to do this. You can deny either (i) or (ii). That is, you must either *deny the implication* and argue that it is possible to have an outcome-related reason not to perform an action of the problematic kind ( $\varphi$ ) even if it is true that the outcome will occur whether you  $\varphi$  or not, or *deny the description* and argue that it is possible to have an outcome-related reason not to  $\varphi$  because the claim that the outcome will occur whether you  $\varphi$  or not is untrue.<sup>7</sup>



Let us first consider ways of denying the implication. In the literature, two main ways of doing this appear. These are the *causal* and *non-causal* responses to the inefficacy argument. The latter category includes quite a rich variety of approaches, such as appeals to fairness, virtue ethics, Kantianism, complicity and membership of a group that harms people. For example, Garrett Cullity (2000) argues that if others do their part of what we all should be doing, it would be unfair of me not to do my part. So, following Cullity, we might argue that it would be unfair of me not to reduce my own greenhouse gas emissions if others do their part in reducing worldwide emissions of greenhouse gases. Dale Jamieson (2007) and Ronald Sandler (2010) argue that the focus on effects of particular acts is misleading, and that we should instead focus on cultivating green virtues. Baylor Johnson (2011) and Christian Baatz (2014) suggest that you have a Kantian “imperfect” duty to reduce your emissions of greenhouse gases, Kutz argues that you might be accountable for environmental harms if you intentionally participate in a way of life that causes such harms, and Derek Parfit (1984) and Anne Schwenkenbecher (2014)

---

<sup>7</sup> Nefsky (2019) helpfully introduces the distinction between responses that deny the implication and responses that deny the description.



try to show that it might be wrong to perform some act if doing so will make your act a member of a set of acts that causes harm. These rejoinders to the inefficacy argument all state, or entail, that you might have a reason to refrain from acting in the relevant way in a collective harm case *even if* it is true that outcome O will occur whether you act in the relevant way or not, and *even if* there is no causal connection between what you do and the relevant outcome.

In contrast, causal responses say that there is a normatively relevant causal connection between acts of the relevant kind and the collective outcome. This is probably the least popular kind of response in the literature, but it does have its advocates. Matthew Braham and Martin Van Hees (2012) argue that Richard Wright's (1985, 2013) NESS condition describes the relevant causal connection. Anton Eriksson (2019) follows David Lewis (1973a, 1986b) in thinking that causation is transitive, and that we must factor in the fragility of outcomes. And Nefsky (2017) argues that you have a reason not to perform the problematic act in a collective harm case if your doing so could be *a non-superfluous part of a cause* of the collective outcome. I will explain each of these positions in due course.

I will present a causal response to the inefficacy argument. I will argue that you have an outcome-related reason not to perform some action  $\varphi$  if and only if (roughly): it is possible that a bad outcome O will occur, possible that O will not occur, and your  $\varphi$ -ing makes O more secure. For instance, you have a climate-change-related reason to refrain from going for a drive in a fossil fuel powered car since it is possible that climate change and its related harms will occur, possible that they will not occur, and going for this drive would make climate change and its related harms more secure. The claim that an act of  $\varphi$ -ing makes an outcome more secure means, also roughly speaking, that fewer things would need to change in order for the outcome to occur given that you  $\varphi$ . In the climate case, fewer other emissions would need to occur in order for a particular climate-change-related harm, such as a flood or drought, to occur. This response is elaborated in Chapter 5, "Reasons for Action", cowritten with Caroline Touborg.

## Denying the Description

Many writers instead deny the description. They dispute that acting in the relevant way has no chance of making a difference for whether the outcome occurs. Much of the discussion here centres on threshold and non-threshold cases. *Threshold cases* are cases where there is some threshold such that if  $n$  acts or fewer of the relevant kind are performed the outcome will not occur, but if  $n + 1$  such acts or more are performed the outcome will occur. In *non-threshold cases*, there is no threshold of this kind. Voting is a typical threshold case. One vote could make a difference to whether some candidate wins or not. For that reason, some have argued, you have

an outcome-related reason to vote. There is always a tiny chance, however minuscule, that whether a certain candidate wins or not will depend on whether you cast your vote. More precisely, it is argued that you have a *subjective* reason to vote given your limited knowledge of the way others will vote. If it is true, unbeknownst to you, that a certain candidate would win whether you cast your vote or not, however, you will lack an *objective* reason to vote. This way, you might have a subjective reason to vote while lacking an objective reason to do so. This is the *expected utility approach*.

If Avram Hiller (2011), Holly Lawford-Smith (2016) and John Broome (2019) are correct, climate change is a threshold case.<sup>8</sup> Building on empirical evidence in one way or another, they argue that a single drive in a fossil fuel powered car has, on average, a negative impact on the climate – and by extension on other people. That is, they argue that your drive in a fossil fuel powered car risks bringing about climate-change-related harms, and that you therefore have a subjective reason not to drive. If they are right about the empirical evidence, the expected utility approach gives the intuitively correct verdict on subjective reasons in this case.

Peter Singer (1980), Alastair Norcross (2004) and Shelly Kagan (2011) take things one step further, and argue that all collective harm cases are threshold cases. To show this, they have to show that there are no non-threshold cases. Kagan urges us to consider a version of Parfit's (1984) famous case of the harmless torturers (here presented in shortened form):

HARMLESS TORTURERS: There are a thousand torturers and one victim. At the start of the day, the victim is already feeling mild pain. Each of the torturers flips a switch, making an instrument affect the victim's pain in a way that is imperceptible. When all the torturers have flipped their switches, the victim is left in excruciating pain.

Here, the relevant outcome is whether the victim is in pain or not. Problematically, it seems that the victim's pain will be the same whether any particular switch is flipped or not. Since the increase in current made by the flipping of a single switch is minuscule, the victim cannot perceive any difference between  $n$  and  $n + 1$  flipped switches. Pain only occurs as a cumulative effect of many flipped switches.

Kagan argues that HARMLESS TORTURERS, contrary to first appearances, is a threshold case. His argument runs as follows. When no switches are flipped, the victim is in no pain. Suppose that it is true that (a) the victim cannot perceive the difference between  $n$  flipped switches and  $n + 1$  flipped switches, as stated in the case. If this is true, it follows that (b) if the victim is in no pain when  $n$  switches are

---

<sup>8</sup> If Broome (2019) is correct, every drive with a fossil fuel powered car makes a difference for which future hurricanes, droughts, floods, etc. that will occur, and for how, when and where they will occur. We cannot, however, know exactly which difference any particular drive will make.

flipped, he will be in no pain if  $n + 1$  switches are flipped. Therefore, the victim will be in no pain when all of the switches are flipped. But this is clearly false. By hypothesis the victim is in extreme pain when all of the switches are flipped. So, our supposition that (a) the victim cannot perceive the difference between  $n$  flipped switches and  $n + 1$  flipped switches must be wrong.<sup>9</sup> Norcross (2004) makes a similar argument. The argument can be generalised to any alleged non-threshold case, and thus it appears to show that there indeed are no such cases.

Many have found this argument wanting. According to Nefsky (2012), it “amounts to giving a sorites argument as though it were a simple reductio proof that there cannot be vague boundaries” (385). That is, if Kagan’s reasoning is correct, we can use similar reckoning to show that a single grain of sand can make the difference between a non-heap and a heap.

While Norcross and Kagan attempt to show that there always is some act that makes a perceptible difference in harm in collective harm cases, Parfit (1984), Barnett (2018) and Broome (2019) take another route. They argue that imperceptible, or even immeasurable, differences might be harms. If they are right, the outcome will not be the same whether or not any particular switch is flipped in HARMLESS TORTURERS. Instead, flipping one switch makes an imperceptible difference for the worse. Since flipping a switch makes an imperceptible difference for the worse (it makes a difference to whether some additional harm will occur or not), each torturer has a suffering-of-the-victim-related reason not to flip his switch. In general, if Parfit and others are correct, you can have a reason to  $\varphi$  (or to refrain from  $\varphi$ -ing) in non-threshold cases, because  $\varphi$ -ing (or refraining from doing so) makes a morally relevant difference to whether some minuscule harm occurs or not.

In the end, it seems to me that some such solution as this will work, but it is not the approach I advocate. Non-threshold cases are not counterexamples to the idea that you have an outcome-related reason to act in a certain way only if your act makes this outcome better or worse. The problem with the expected utility approach – and more broadly with the idea that you only have an outcome-related reason to act in a certain way if your act makes this outcome better or worse – instead shows in cases of overdetermination. Consider, for instance, the following case:

ASSASSINS: Two assassins simultaneously shoot a victim, and do so independently of each other. Each shot pierces the victim’s heart and so was sufficient for the death of the victim. The victim dies.<sup>10</sup>

---

<sup>9</sup> Here I have simplified Kagan’s argument. In Chapter 8, I go through it more thoroughly.

<sup>10</sup> Discussion of this case is commonplace in the literature (see for instance Parfit 1984; Fischer & Ravizza 1998; Brahm & Van Hees 2012).

Intuitively, it seems that each assassin has an outcome-related reason not to shoot the victim (I assume that the death of the victim is a bad thing). Still, the victim will die whether or not any particular assassin takes a shot. So, how can we explain the intuition that each assassin has a reason not to shoot?

The expected utility theorist (Singer, Norcross, Kagan, etc.) typically argues that each assassin does have a reason not to shoot, since *as far as they know* their shot *could* make a difference to whether the victim dies. That is, each assassin has a subjective (but not an objective) reason not to shoot. I think this verdict is mistaken, and that the assassins also have an objective reason not to shoot. However, this point is hard to defend. Our intuitions about objective reasons in cases like ASSASSINS might go either way.

When we instead ask who is blameworthy for killing the victim, our intuitions are clearer. It seems that both assassins deserve to be blamed for the death of the victim. Still, if we hold on to the idea that the only thing that matters morally is whether the occurrence of an outcome is dependent on your  $\phi$ -ing, we must conclude that neither assassin can be blamed for the victim's death. After all, it is true of each assassin that the victim would have died whether or not this assassin had shot the victim. So, on this account, it is true of each assassin that we cannot blame him for the death of the victim. We can only blame him for attempted murder, or more precisely for shooting even though he had a subjective reason not to. This surely is an unattractive conclusion.

This brings me to the second aim of this thesis, which is to give an account of the conditions under which you are blameworthy for an outcome – an account that gives the right verdict also in collective harm cases. This is the topic of Part Two of this thesis. Briefly, building on the idea that you have an outcome-related reason to refrain from  $\phi$ -ing if your doing so will increase the security of the harmful outcome, I argue – again together with Touborg – that you are blameworthy for some harmful outcome O if and only if your having a poor quality of will towards O is a cause of this outcome. Here, I take it that C is a cause of E if and only if C makes E more secure *and* C is process-connected to E.

This will be explained in detail in due course. However, the general idea is that you cause an outcome if what you do makes this outcome closer to happening (or further from not happening) *and* if this outcome is connected to what you do in the right way. What each assassin does, for instance, makes the outcome more secure. Each shot moved the death of the victim further from not happening. Additionally, what each assassin does is process-connected to the death of the victim. There is a process leading from the firing of each gun, via the bullet's flight through the air and its piercing of the victim's heart, to the death of the victim. So, on the account I present with Touborg, what each assassin does is a cause of the death of the victim.

Still, we do not say that you are blameworthy for a harmful outcome simply because what you did was a cause of this outcome. That inference is blatantly unsafe. You might have been coerced into acting as you did, or you maybe you did not know that your act would have harmful consequences. This is why we say that you are blameworthy for some harmful outcome  $O$  if and only if your *poor quality of will towards  $O$*  caused this outcome. Thus, in ASSASSINS, each assassin is blameworthy for the victim's death because, first, his blatant disregard for the victim increased the security of the victim's death – had he cared as required about the victim, and not fired his gun, and the victim's death would have been less secure – and second, his blatant disregard for the victim is connected to the victim's death (via his firing his gun, the bullet's flight through the air, and so on).

The main advantage of our accounts of outcome-related reasons and when you are blameworthy for an outcome is their ability to provide intuitively correct verdicts across a large range of cases. They give correct verdicts not only in collective harm cases of the threshold and non-threshold varieties, but also in a range of cases I have not yet introduced, including what are known as pre-emption cases, switching cases, Frankfurt-style cases, and more. This will also become evident as we go along.

## About This Thesis

The title of this thesis is *Reasons, Blame, and Collective Harms*. The thesis has two parts: one about reasons and one about blame. Both concern collective harms. In simple terms, in Part One, I argue that you have an outcome-related reason to  $\varphi$  if and only if  $\varphi$ -ing increases the security of a positive outcome. In Part Two, I argue that you are blameworthy for  $X$ , where  $X$  is some action, omission or outcome, if and only if your bad quality of will is a non-deviant cause of  $X$ . I show that these principles explain our intuitions about reasons and blameworthiness in collective harm cases, but also in a wide range of other cases.

To be more specific, this is what the thesis is about:

In Chapter 2, I argue that we need a causal solution to the inefficacy problem. Unless we can show that there is a relevant causal connection between  $\varphi$ -ing and the outcome in collective impact cases, appeals to fairness, virtue ethics, Kantianism, complicity, reasons to take collective action, and membership of a group will inevitably be unsuccessful. For instance, you might think that it would be unfair if I do not pull my weight in what we all ought to be doing, or that it would be unfair if I do not try to counteract climate change when others do. The problem is that this idea cannot explain why I have a climate-change-related reason to refrain from joy-guzzling – not unless we can show that refraining from joy-guzzling counts as counteracting climate change and its related harms. If whether some climate-change-related harm occurs when I joy-guzzle is all that matters, it seems that

refraining from joy-guzzling cannot pull any weight at all in the debate over what we all should be doing to counteract climate change. Sinnott-Armstrong (2005), Nefsky (2015) and others have made this point before.

In Chapter 3, I consider Braham and Van Hees' (2012) and Eriksson's (2019) causal responses to the inefficacy argument. I argue that they fail because they do not capture the idea that the relevant causal connection is one of contribution. I leave it open what a causal contribution is. It could be something that raises the probability of the outcome, raises the security of the outcome, makes the outcome happen sooner rather than later, makes the outcome occur rather than not, or some other dynamic feature of this sort. Just to give a flavour of the idea, Braham and Van Hees rely on the NESS condition of causation in their analysis. However, this account of causation gives the wrong verdict in switching cases like the following:

THE ENGINEER: an engineer is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch, so that the train travels down the right-hand track, instead of the left. Since the tracks reconverge up ahead, the train arrives at its destination all the same; let us further suppose that the time and manner of its arrival are exactly as they would have been, had she not flipped the switch.

(Hall 2000: 205)<sup>11</sup>

Suppose further that the train arrived late at the station. Did the engineer's flipping her switch cause the train to arrive late at the station? Intuitively, it did not. However, NESS entails that it did. According to NESS, an action is a cause of an outcome if it was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the occurrence of the outcome. And, in TRAIN SWITCH, the engineer's flipping of the switch was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the train's late arrival. Had she not flipped the switch, this set would not have been sufficient for the occurrence of the outcome (while another set would: the set that was also sufficient for the train's remaining on the left track). Further, if we combine NESS with an account of outcome-related reasons, or an account of blameworthiness and praiseworthiness, we obtain the result that the engineer might have had train-arriving-late-related reasons not to flip the switch, and that she might deserve blame for the train's late arrival if she does. This seems strange. On my analysis, it seems strange precisely because flipping the switch does not contribute to the train's late arrival.

---

<sup>11</sup> Switching cases like this are common both in the causation literature and in the literature on moral responsibility. For the former, see e.g. Hall (2000, 2007), and Paul and Hall (2013). For the latter, see e.g. Foot (1967), Thomson (1976), Van Inwagen (1978) and Fischer and Ravizza (1998).

Chapter 4 considers what might be the best attempt at solving the inefficacy problem so far, namely Nefsky's (2019). Nefsky's view is that you have a reason not to  $\varphi$  if  $\varphi$ -ing could be a *non-superfluous part of a cause* of the harmful collective outcome. I offer some counterexamples to this idea, but the main take-away is that Nefsky's appeal to non-superfluity could be redundant. If we are more careful when spelling out the relevant causal connection – which I think involves contribution (and more specifically, security-raising), not being part of a cause – it turns out that we have no need for an analysis of non-superfluity in our account of outcome-related reasons.

In Chapter 5 “Reasons for Action”, Touborg and I present our positive proposal of outcome-related reasons. As already indicated, we suggest that you have an outcome-related reason to refrain from  $\varphi$ -ing just in case the outcome is negative and  $\varphi$ -ing makes it more secure.<sup>12</sup> We spell this out in terms of possible worlds. In HARMLESS TORTURERS, for instance, we would say that each torturer has a reason not to flip his switch since doing so would reduce the distance between the actual world and the closest world where the victim is in extreme pain. This invites the objection that it is unclear why flipping a switch would make the victim's being in extreme pain more secure. Just as you cannot decide the exact distance between a point and an interval with fuzzy boundaries, you cannot decide the distance between the actual world and the closest world where the victim is in extreme pain – or so it might be argued. We respond that there is an intuitive sense in which the flipping of each switch takes us closer to the harmful outcome. We also argue that contemporary theories of vagueness support this conclusion.

In Chapter 6, I consider further examples that were not discussed in “Reasons for Action”.

It can be seen, then, that Chapters 2–6 focus on attempts to refute the first premise of the inefficacy argument. In other words, they consider the potential to “deny the implication”. Chapters 7–9, by contrast, look at refutation of the second premise, asking whether we can “deny the description”. In Chapter 7, I discuss empirical arguments against the inefficacy argument. I argue that appealing to empirical evidence is successful as long as we can trust the expected utility approach, but the mere possibility of non-threshold cases should make us hesitate to accept this approach.

In Chapter 8, I argue that Kagan's argument that there are no non-threshold cases is inconclusive. Others have argued this before. In particular, Nefsky (2012) argues that Kagan essentially provides a sorites argument as though it were a reductio of the idea that there can be vague boundaries. I pick up where Nefsky left off. I argue

---

<sup>12</sup> Likewise you have an outcome-related reason to  $\varphi$  just in case the outcome is positive and  $\varphi$ -ing makes this outcome more secure. On the full analysis, it is not necessary to separate the positive and the negative cases.

that our best theories of vagueness (the epistemic view of vagueness, a three-valued logic and supervaluationism) all entail that there is a threshold in collective harm cases generally. However, the analysis points to another problem with Kagan's conclusion: the thresholds are conceptual and not necessarily perceptible. Given that only perceptible differences matter morally, passing such a threshold does not necessarily trigger morally relevant harm, contrary to what Kagan claims. I also discuss a number of interpretations of Kagan's argument, and find them wanting.

Chapter 9 also concerns the idea that the mere possibility of non-threshold cases poses a problem for the expected utility approach. I argue that there is a successful argument that unlike Kagan's shows that non-threshold cases are impossible. If this is correct, non-threshold cases are not counterexamples to the expected utility approach. However, I argue that the expected utility approach faces another problem. It fails to properly explain our objective reasons in threshold cases. This is more obvious when we consider blameworthiness than it is when we look at outcome-related reasons. This concludes the first part of the thesis.

Part Two is shorter. It is about blameworthiness. More precisely, it is about the circumstances under which others are warranted in directing negative reactive attitudes like resentment and indignation towards you. This part of the thesis opens with a chapter in which I distinguish what I call *blameworthiness* from *blameworthiness for*. Assessments of *blameworthiness* are evaluations of the character, the moral fibre, or the quality of will, of a person (perhaps in combination with some further condition), whereas assessments of who is *blameworthy for* something ask whose fault something is (if it is anyone's fault). In the former, causation plays no role. In the latter, causation is essential.

The account Touborg and I propose is an account of *blameworthiness for*.<sup>13</sup> Zimmerman (2002) and others have argued that questions about what you are blameworthy for are otiose; that is, that they do not affect how blameworthy you are. I disagree. I argue that we might be warranted in reacting more negatively to someone who caused some harm out of a lack of care than we do to someone who likewise did not care, but, as a result of luck, did not cause any harm. In other words, I argue that there is such a thing as resultant moral luck, at least when it comes to *blameworthiness for*. This is a controversial position. Nevertheless, I think it is correct. Even if I turn out to be mistaken on this point, assessments of what you are blameworthy for are not otiose. They matter in many contexts – for instance, in the settlement of legal disputes, and when we are debating whose fault something is.

In the central chapter of Part Two, Chapter 11, "You Just Didn't Care Enough", Touborg and I present our account of the conditions under which you are blameworthy for something. The core thought here is that you are blameworthy for

---

<sup>13</sup> Note that on this distinction you might be blameworthy *for* being a bad person if your previous choices in some relevant way caused you to be such a person.



an action, omission or outcome if some substandard quality of will of yours non-deviantly caused this action, omission or outcome.

This idea is elaborated and refined in a number of respects in Chapters 11–13. In Chapter 12, I go through some of the finer details behind the idea that a cause must be process-connected to its effect, and consider what it is for a cause to *non-deviantly* cause something. In Chapter 13, I consider some further cases, and try to show that the account Touborg and I offer gives the intuitively correct verdicts also in these cases. For instance, I argue that our account delivers the right verdict about a case called “Pinned-In Sharks”, which is a case that poses trouble for John Martin Fischer and Mark Ravizza’s influential account of moral responsibility. In addition, I defuse a potential counterexample.

Chapter 14 is the final chapter of the thesis. In it, I discuss Carolina Sartorio’s (2004, 2015, 2016) arguments for the view that you might be blameworthy for an outcome without causing it. If she is right, I am wrong. I argue that her arguments are unconvincing, but in an interesting way. In the course of doing this, I have an opportunity to discuss James A. McLaughlin’s (1925-26) much-debated example of the thirsty traveller. The traveller dies of thirst in the desert after A replaces the water in his canteen with sand, and B steals the canteen unaware that it is filled with sand (Mackie 1974; see e.g. Gavison, Margalit, & Ullmann-Margalit 1980; Hart & Honoré 1985; Wright 2013; Talbert 2015; Bernstein 2019). In the course of this discussion, I draw upon all of the resources of our account of *blameworthiness for*. Among other things, I consider what the relevant possibility horizon is, and what it is to non-deviantly cause an outcome.

## 2. Non-Causal Responses

Some argue that you might have an outcome-related reason to  $\phi$  even if there is no causal connection between  $\phi$ -ing and the relevant outcome. They may appeal to fairness, virtue ethics, Kantianism, complicity, reasons to take collective action or membership of a group that causes harm. I will argue that these approaches succeed only if we also accept that there is some causal connection between  $\phi$ -ing and harm. If I am right, it does not follow that virtue ethics, Kantianism, etc. are mistaken. It means merely that virtue ethics, Kantianism etc. cannot stand on their own as responses to the inefficacy argument. They will give intuitively correct verdicts in some important collective impact cases only if they are complemented with an appropriate account of causation, or in some cases with an appropriate account of outcome-related reasons based on causation. Many of the points I make in this chapter have been made before.<sup>1</sup> My contribution here is not so much to give new arguments as to give a comprehensive overview of the issue, and to elaborate some previously given arguments.

I will follow Julia Nefsky (2015, 2021) in arguing that non-causal solutions to the inefficacy problem run into either the disconnect problem or the superfluity problem (or in the worst case scenario, both). They run into *the disconnect problem* when “the reason for action identified is not a reason that can count as addressing what is at issue in the problem of collective impact” (2021). For instance, some might argue that a virtuous person would not joy-guzzle, and hence that virtue ethics might help us avoid the inefficacy problem. However, if they next say that the virtuous person would not go joy-guzzling because doing so is a waste of time, and because there are better things to do, such as visiting a friend or contemplating life, they have not really addressed the problem at issue. They have failed to cite a climate-change-related reason explaining why the virtuous person would refrain from joy-guzzling. The reasons they have given are of a different kind.

Characteristically, non-causal responses that are disconnected from the collective impact fail to reliably distinguish relevant outcome-related reasons. This is true of the fairness approach. Thus, think of a world where no one is reducing their greenhouse gas emissions with the result that climate change and its related harms are looming on the horizon. You might think that people in this world have a

---

<sup>1</sup> For instance, by Parfit (1984), Sinnott-Armstrong (2005), Sandberg (2011) and Nefsky (2015, 2019, 2021).

climate-change-related reason to reduce their emissions. However, you could never explain this by appealing to fairness. The appeal to fairness essentially says that it would be unfair of you not to pull your weight in a shared enterprise we should all be involved in, when others are pulling their weight. However, in this world, no one else is pulling their weight, so it would not be unfair of you not to pull yours.

Somewhat differently, non-causal responses run into *the superfluity problem* if the identified reason for action does not apply unless we assume that there is some relevant causal connection between the act and the outcome. For instance, someone might cite climate change as the reason why a virtuous person would not go joy-guzzling. However, it is unclear why a virtuous person would refrain from joy-guzzling for climate-change-related reasons if there really were no relevant causal connection between joy-guzzling and climate change.

Whether a non-causal response runs into the disconnect problem or the superfluity problem will depend on the specifics of it. Some Kantian responses, for instance, may run into the disconnect problem, while others run into the superfluity problem.

In sum, if a particular non-causal account runs into the disconnect problem, it gives counterintuitive verdicts in some important cases, and if it runs into the superfluity problem, it gives intuitively correct verdicts but only because it presupposes that there is a relevant causal connection between  $\phi$ -ing and the collective outcome. These points will become clearer as we proceed.

## Fairness

Fairness is often invoked in free rider cases. The literature on this issue is huge. Seminal works include H. L. A. Hart's (1955) "Are There any Natural Rights?", Mancur Olson's (1965) *The Logic of Collective Action* and John Rawls' (1971) *Theory of Justice*. In free rider cases, the relevant agents who might or might not  $\phi$  are also those who benefit from the outcome. Take mass transport as an example. The existence of efficient and affordable buses is a common good. There is a group that benefits from this system. Still, anyone in this group might think in the following way: Why should I pay my bus fare if I can sneak on without doing so? The payment or non-payment of €3 will surely make no difference to the future availability of affordable buses. A common answer is: I am only able to gain the benefits of there being efficient and affordable buses because others pay their bus fares. So, not paying my fare would be unfair. In general, when we benefit from some common good, it is unfair if some but not others contribute to this common good. This unfairness grounds a reason to contribute to the common good.

Fairness can also be invoked in collective impact cases. Consider, for instance, the following case, first described by Derek Parfit (1984):

[DROPS OF WATER:] Imagine that there are ten thousand men in the desert, suffering from intensely painful thirst. We are a group of ten thousand people near the desert, and each of us has a pint of water. We can't go into the desert ourselves, but what we can do is pour our pints into a water cart. The cart will be driven into the desert, and any water in it will be evenly distributed amongst the men.

If we pour in our pints, the men's suffering will be relieved. The problem is, though, that while together these acts would do a lot of good, it does not seem that any individual such act will make a difference. If one pours in one's pint, this will only enable each man to drink an extra *ten thousandth* of a pint of water. This is no more than a single drop, and a single drop more or less is too minuscule an amount to make any difference to how they feel.

(Nefsky 2017: 2743-44)<sup>2</sup>

Here, we might think that we – the pint holders – have a collective obligation to alleviate the suffering of those suffering from thirst in the desert. We might also think that if others contributed their pints, it would be unfair of me not to contribute mine. That is, we might think that if others contribute their pints, I have a reason to contribute mine. This seems to be an outcome-related reason. Could this kind of reason explain the reasons intuition?

## **Making Others Work Harder**

To answer this question, we must ask what unfairness is, more exactly. Four main types of unfairness can be discerned in the literature. First, it might be unfair if others have to work harder if I do not do my part. For instance, Jonathan Glover (1975) considers a case in which a car needs to be pushed up a hill. We are eight people who could do this, but it only takes six of us to accomplish the task. Here, Glover argues that it would be unfair if you and I just sat and watched while the other six push the car up the hill, and that this is because they have to push harder if we do not help.<sup>3</sup> To take another example, it seems unfair if you always do the dishes while I just sit and relax (unless I have some justification for doing so). Since I never do the dishes, you have to work more.

As Glover (1975), Parfit (1984) and others have pointed out, this kind of unfairness is irrelevant in collective impact cases. Here, my contribution does not make any difference to whether others have to contribute more. As Glover puts it in relation to voting, “no-one has to vote harder because I do not vote” (182). Similarly, it is not as if the other pint holders in DROPS OF WATER have to donate more if I do not donate my pint. By hypothesis, no one will suffer more if I do not donate my pint. So, since we are currently discussing the inefficacy argument and collective impact cases, we can set this kind of unfairness aside

---

<sup>2</sup> I use Nefsky's (2017) version of the example because it brings out the problem even more clearly.

<sup>3</sup> Glover is commenting on an argument presented by Lyons (1965).

## Benefitting Without Doing One's Part

Second, it might be unfair if I benefit from the collective outcome brought about by enough people's  $\varphi$ -ing while not  $\varphi$ -ing myself. To gain the benefits of there being efficient and affordable buses because others pay their bus fares while not paying myself might be one example of this. To enjoy a clean home because my spouse or roommate cleans it while not doing my part of the cleaning might be another. Perhaps driving to work without running into a traffic jam because others are using mass public transport is a third. Jason Brennan (2009) suggests something along these lines in relation to climate change.

*We* should pollute less because pollution harms us all, but *I* should pollute less because, all things equal, it is unfair for me to benefit from polluting as I please while others suffer the burden of polluting less.

(Brennan 2009: 541)<sup>4</sup>

Hart's (1955) account of fairness is of this broad sort as well:

when a number of persons conduct any joint enterprise according to rules and thus restrict their liberty, those who have submitted to these restrictions when required have a right to a similar submission from those who have benefited by their submission (185).

(Hart 1955: 185)

Rawls (1971) has a similar formulation.

On this understanding of fairness, the problem is not that others have to work harder when I do not do my part (though this may be the case), but rather that I benefit from the collective outcome brought about by enough people's doing their part while not doing my part in bringing that outcome about.

This account of fairness faces counterexamples. I might lack a reason to do my part in bringing about some collective outcome even if I happen to benefit from this outcome. As Nozick argues, you are not entitled to "decide to give me something, for example a book, and then grab money from me to pay for it, even if I have nothing better to spend the money on" (1974: 95). Here, I will assume that there is

---

<sup>4</sup> Brennan's main argument concerns voting, but the idea is the same in both cases. Brennan states his idea somewhat differently at different points in his paper. In some places the idea is instead that those who have caused climate change have an obligation to clean up: fairness requires everyone who has caused climate change to help out by doing this.

a modified account that avoids Nozick's objection,<sup>5</sup> and instead concentrates on another issue.

The idea that it is unfair to benefit from the collective outcome without doing one's part does not explain the reasons intuition in all collective impact cases. This is simply because there are collective impact cases where you do not benefit from the collective outcome. In DROPS OF WATER, for instance, you are not one of those benefitting from the collective outcome. You are not one of those who will have their suffering alleviated if everyone donates their pint. So, in this case, we cannot explain the intuition that you have reason to donate your pint on the basis that, if you do not, you will benefit from the collective outcome without doing your part. The same could be said about any collective impact case in which you are not one of those who benefits from the collective outcome.<sup>6</sup>

### **Doing One's Part**

Cases like DROPS OF WATER are cases where, while you do not benefit from the collective outcome, you (and others) have an obligation to bring it about. In these cases, you might instead think along the following lines: If collective outcome  $O$  will occur if enough of us  $\phi$ , and if we have a collective obligation to bring about  $O$ , it will be unfair of me not to  $\phi$  if enough others  $\phi$ . For instance, it would be unfair of me not to donate my pint in DROPS OF WATER when others are donating theirs, because the suffering of the people in the desert will be relieved if enough of us donate our pints and we have a collective obligation to bring about the alleviation of suffering. Similarly, if we should reduce our greenhouse gas emissions in order to limit further climate change and its related harms, and if others are going the extra mile in order to reduce their emissions, you might think that it would be unfair of me to continue driving my gas-guzzling car just for fun.

In order to establish the present understanding of unfairness – i.e. that it is unfair, in all collective impact cases, not to do one's part in satisfying a collective obligation if others are doing their part – we need to establish that the participants in these cases have a collective obligation to bring about the relevant outcome. To establish this in a principled way requires some work. Here, for the sake of argument, I will

---

<sup>5</sup> Arneson (1982) suggests such an account.

<sup>6</sup> Things become more complicated if we factor in that I prefer the suffering of the people in the desert to be alleviated. If I have this preference, I will benefit if their suffering is alleviated: a preference of mine will now be satisfied. Therefore, it might be unfair of me not to donate my pint, according to the understanding of fairness currently under consideration. I will benefit from the collective outcome, brought about when enough people donate their pints, without donating mine. Still, I think we can set this complication aside. Fairness is not contingent on preferences in this way. It would be unfair of me not to donate my pint if others donate theirs whether or not I prefer the suffering of the people in the desert to be alleviated.

simply stipulate that the participants have such obligations in the collective impact cases under consideration. With this stipulation in place, the understanding of fairness in which we are interested applies to the collective impact cases under consideration.

On this understanding of fairness, the problem is not that others have to work harder if I do not do my part, or that I benefit from the collective outcome while not doing my part in bringing it about. The problem is simply that it would be unfair for me to keep my pint while others are donating theirs if we have a collective obligation to fill the cart, or for me to go joy-guzzling while others are bearing the burden of reducing their emissions if we have a collective obligation to reduce emissions. It is a matter of equality.

This understanding of fairness invites the *levelling down objection* (as indeed did the previous one). How can a change be an improvement if it merely consists in the better-off losing some benefit? This objection may need some clarification. We might think that equality is instrumentally good. For instance, if you and I are thirsty, and, as chance would have it, I have plenty of water while you have none, it seems that things would be better if I gave you some of my water. That way, you also could quench your thirst. Here, things are better if they are more equal. However, in DROPS OF WATER, no one is better off just because I donate my pint. Likewise, when it comes to climate change, no one will be better off just because I take the bus to work instead of the car. In these cases, things are just as bad for everyone else if I do not do my part in bringing about the collective outcome.<sup>7</sup> Still, doing my part is a cost for me. I lose a pint of water, and have to get to work in a less convenient way. We might then ask: How could a change resulting in a situation which is better for no one, but worse for someone, be an improvement? How can I make things better by giving up my pint if this makes no one better off? Likewise, how can I make things better by taking the bus to work instead of the car if no one becomes better off as a result? To believe that I can do this is to believe that equality is intrinsically valuable. Glover (1975) calls such a belief “a Dog-in-the-manger version of justice” (182).<sup>8</sup>

In contrast with Glover and others, we might think that equality has some value in itself. However, as Nefsky (2015) argues, there is a deeper issue lurking here: Even if we accept that equality has value in itself, and as a consequence that I have a reason in DROPS OF WATER to level down by giving up my pint, this does not show

---

<sup>7</sup> At least, this is true if climate change is a non-threshold case. As Broome (2019) argues, empirical evidence indicates that climate change is not such a case (see discussion in Chapter 7). However, for the sake of discussion and illustration, I follow others in thinking about climate change as a non-threshold case throughout this chapter. You might find this problematic, but I hope that you look past this difficulty and still get the gist of the argument.

<sup>8</sup> For discussion of the levelling down objection to egalitarianism see, for instance, Parfit (1997), Holtug (1998) and Temkin (2000).

that I have an alleviation-of-suffering-related reason to give up my pint. At most, it shows that I have an equality-related reason to do so. Similarly, even if we accept that equality has some value in itself, and as a consequence that I have a reason to reduce my emissions, this does not show that I have a climate-change-related reason to do so. At most, it shows that I have an equality-related reason to do so. Problematically, collective-outcome-related reasons and equality-related reasons do not always overlap, which means that we will run into the disconnect problem. This shows in three different ways.

For one thing, we sometimes have a collective-outcome-related reason but no equality-related reason. Consider a scenario similar to DROPS OF WATER but where no one donates their pint. Here, you might still think that each of us had a reason to do donate our pint. After all, there is a possibility that the suffering of the people in the desert will be alleviated if just enough of us donate our pints, and the alleviation of suffering would be a good thing. However, the appeal to equality cannot explain this. It only kicks in if others donate their pints. Likewise, think of a scenario where scientists have just discovered that massive emissions of carbon dioxide lead to climate change and its related harms, but where no one reduces their emissions. You might think that people in this scenario have a climate-change-related reason to reduce their CO<sub>2</sub> emissions. However, we could never explain this by appealing to equality. Again, this appeal only gains traction if others reduce their emissions.

For another thing, we sometimes have an equality-related reason in the absence of a collective-outcome-related reason. We might think that you have no alleviation-of-suffering-related reason to donate your pint to a cart that already is full. To do so would just result in an overflow. At best, it would be a completely superfluous thing to do. However, this is not the verdict we obtain if we take seriously the idea that you have a reason to donate your pint if others have donated theirs. Doing that, we have to conclude that you have a reason to donate your pint even if the cart already is full. You have such a reason since this will make things more equal. (Here, I am assuming that the cart is full because the 10,000 others have already donated their pints).

Finally, as Nefsky (2015) points out, if equality is what matters, it is unclear why I specifically have a reason to pour my pint into the cart, instead of, say, emptying it on to the ground. Either way, I end up having as much water as the other pint holders. Again, we see that equality-based reasons and outcome-related reasons come apart.

At this point, it might be suggested that we have described the argument from fairness in the wrong way. What matters, it might be said, is not that it will be unfair if I do not *donate my pint* when others donate theirs, but that it will be unfair if I do not *contribute to the alleviation of suffering* if others do so. This brings us to the fourth sense of fairness discussed in the literature.



## Pulling One's Weight

Fourth and finally, some will argue that the following is what is unfair: my not contributing to what we all should be doing while others do. More precisely, they will suggest that if outcome O will occur if enough of us contribute to O, and if we have a collective obligation to bring about O, it would be unfair of me not to contribute to O if a sufficient number of others are doing so. This is essentially Garrett Cullity's explanation of why each of us has a reason to act in the relevant way in collective impact cases. He argues that I should pour my pint into the cart in DROPS OF WATER because, if I do not, I will "be relying on others to do what we ought collectively to be doing, without contributing [myself]" (2000: 15). As he also puts it, "It is a matter of pulling my weight in what we all ought to be doing" (2000: 17). Cullity (2019) applies this argument to climate change and some other real-life challenges.

Like fairness-as-equality, this kind of fairness applies to collective impact cases. The problem is not, for instance, that others have to work harder if I do not do my part, or that I benefit from some outcome without doing my part in bringing it about, but that it would be unfair of me not to pull my weight when others are pulling theirs.

As Cullity (2000) points out, if we appeal to the idea that it will be unfair if I do not contribute when others are doing so, we can explain why you do not have a reason to pour your pint into a cart that already is full. To do so would not be to contribute to compliance with the collective imperative of alleviating suffering. In a similar vein, we can now explain why you have a reason to pour your pint into the cart rather than onto the ground. Pouring it into the cart does contribute to compliance with the collective imperative (given that the cart is not already full), whereas pouring it onto the ground does not. So, by saying that it is unfair not to *contribute* while others do rather than saying that it is unfair not to *do one's part* while others do, we can avoid an important objection to the fairness approach.

However, as Nefsky (2015) points out, Cullity's solution runs into the superfluity problem. The claim that you have a reason to contribute your pint if others contribute theirs straightforwardly presupposes that acting in the relevant way *contributes* to the outcome. Unless we can show that pouring a pint into the cart contributes to the alleviation of suffering, Cullity's solution will not get off the ground. The same could be said in the case of climate change. Nefsky puts this point in the following way, "If acting in the relevant way won't make any difference then it does not seem that it pulls *any weight at all*" (2015: 257). This seems correct. Acting in the relevant way must make some kind of difference to the outcome in order to count as a contribution. It must raise the probability of the outcome, raise the security of the outcome, make the outcome occur sooner rather than later, make it occur at all rather than fail to occur, be a non-superfluous part of the cause of compliance with a collective obligation, or something of this kind.

As a final point, Cullity's solution does not completely escape the disconnect problem. It still cannot explain the intuition that each of us has a reason to donate our pint in DROPS OF WATER even if no one else does so, or the intuition that each of us has a reason to reduce our emissions even if no one else does so. Cullity (2000) acknowledges this, saying that he has concentrated on explaining "one way in which individual imperatives can be derived from collective ones. Perhaps there are others" (2000: note 22). I think we can do better than this. If we accept that donating a pint in some relevant sense contributes to the alleviation of suffering (something that Cullity is obliged to accept anyway, since his solution presupposes that donating a pint contributes to this outcome), we can argue that each of us has an alleviation-of-suffering-related reason to donate our pint. Similarly, we can argue that each of us has a climate-change-related reason to reduce our emissions. If we do, it will not matter whether others contribute. What will matter is that there is a possibility of reaching the beneficial outcome, and that each contribution takes us closer to reaching it.

## Virtue Ethics

Addressing the inefficacy problem, especially when it comes to environmental harms, some have argued that we should focus on virtuous traits of character, not the effects of particular acts. Dale Jamieson (2007) argues that utilitarians should become virtue theorists when dealing with the problem of climate change. Since utilitarians think that you are morally required to act in such a way as to produce the best outcomes, and since, in this case, a focus on character traits will produce better outcomes than a focus on outcomes of particular actions, utilitarians should focus on the traits. As Jamieson puts it:

Instead of looking to moral mathematics for practical solutions to large-scale collective action problems, we should focus instead on non-calculative generators of behavior: character traits, dispositions, emotions and what I shall call "virtues".

(Jamieson 2007: 167)

Ronald Sandler (2010) similarly argues that virtue-oriented ethical theories provide a compelling response to the inefficacy argument (or "the problem of inconsequentialism", as he calls it). Among other things, such theories do not evaluate discrete actions entirely on the basis of their outcomes. Their evaluations are also constructed "on the basis of patterns of behavior or activities throughout a person's life, as well as patterns among people or communities" (176). According to Sandler, these theories can be stated generically as follows:

THEORY OF VIRTUE: A character trait is a virtue to the extent that its possession is generally conducive to promoting the good; and a character trait is a vice to the extent that it is generally detrimental to promoting the good.

PRINCIPLE OF RIGHT ACTION: An action is right to the extent that it is virtuous.

(Sandler 2010: 176)

As Sandler points out, these principles are underspecified. THEORY OF VIRTUE is silent on the nature of the good. Commonly, virtues are defined in terms of their relationship to *eudaimonia*, which can be roughly translated as “the good life”, “true happiness” or “the kind of happiness worth seeking and having”. According to eudaimonist virtue ethics, being virtuous typically helps you live in true happiness (see Rosalind Hursthouse 1999; Julia Annas 2011). In contrast, Julia Driver (2001) argues that a moral virtue is “a character trait that systematically produces or give rise to the good” (108), where “the good” is to be understood in a more consequentialist fashion so that the consequences for everyone count. Jamieson (2007) follows Driver on this point. THEORY OF VIRTUE subsumes both approaches.

Here, I will be agnostic about the nature of the good. For the sake of argument, I will, however, assume that the good is such that climate-change-related harms are considered bad. If, like Hursthouse and others, we understand the good in eudaimonist terms, we will be in a position to argue that climate-change-related harms are detrimental to my true happiness. Alternatively, if, like Driver and others, we understand the good in a more consequentialist fashion, so that consequences for everyone’s true happiness count, we can argue that climate-change-related harms are detrimental to people’s true happiness. I will also assume that the nature of the good is such that the alleviation of suffering of the people in the desert is good. This seems straightforwardly true if we understand the nature of the good as Driver does. But the idea will need further explaining if we instead interpret the good in eudaimonist terms. Why would alleviation of the suffering of the people in the desert affect my true happiness? One possible answer is that a virtuous person would want the suffering of those in desert to end, and therefore the alleviation of their suffering would be part of my true happiness (whether or not it actually makes me happy). Perhaps there are other possible answers as well.

The PRINCIPLE OF RIGHT ACTION is also underspecified. It does not say anything about what it is for an act to be virtuous. An action could be virtuous if it is what an agent with virtuous character traits would do (as Hursthouse, 1991, argues), if it is what a virtuous agent would advise one to do in the circumstances, or if it is what a fully virtuous agent (a *phronimos*) would do in those circumstances.

## What Would the Fully Virtuous Person Do?

A virtue-oriented theory like the one Sandler advocates does not help us escape the inefficacy problem. Consider first the PRINCIPLE OF RIGHT ACTION, and ask why it would be virtuous not to joy-guzzle. Perhaps refraining from joy-guzzling is the virtuous thing to do because this is what a fully virtuous agent – a *phronimos* – would do. Here, I take it that the fully virtuous agent is someone who does the right things for the right reasons. She realises what the morally relevant considerations are in any situation, weighs them properly, and acts accordingly. If this is so, and if there truly are no climate-change-related reasons to refrain from joy-guzzling (which is the case if the inefficacy argument is correct), the fully virtuous agent would realise this. She would not see climate-change-reasons to refrain from joy-guzzling if there are none. Rather, she would see that her joy-guzzling has no morally relevant impact – it will affect neither her true happiness nor the true happiness of others.

She might still refrain from joy-guzzling, but if she does, she will do so for other reasons. In fact, it is quite reasonable to think that the fully virtuous agent would do something other than joy-guzzling. She would probably visit a friend, finish an important paper, spend time in nature, contemplate life, or something like that, instead of seeking cheap thrills in a gas-guzzling sport utility vehicle. But surely, the fully virtuous agent does not refrain from joy-guzzling for reasons that do not exist.

The point here is not to say that a fully virtuous agent would not see a climate-change-related reason not to joy-guzzle. She agent would most likely see such a reason. Rather, the point is that *if the inefficacy argument holds*, the fully virtuous person lacks a climate-change-related reason not to joy-guzzle, and would realise this. So, unless we can show that the fully virtuous agent has a reason not to joy-guzzle – that is, unless we can find some other solution to the inefficacy problem – we have no good grounds for thinking that the fully virtuous person would refrain from joy-guzzling for climate-change-related reasons. That is, in trying to explain the reasons intuition by appealing to what the fully virtuous person would do, we run into the superfluity problem.<sup>9</sup>

I should add a caveat here. Perhaps one single drive in a fossil fuel car *does* make a difference to which climate-change-related harms will occur. If John Broome (2019) is correct, because of the atmosphere’s instability, “we should expect global weather in a few decades’ time to be entirely different if you go joy-guzzling on Sunday from what it would have been had you stayed at home” (113). Further, although we cannot know exactly how my leisure drive will affect the climate, we have reasons

---

<sup>9</sup> We also run into the disconnect problem if we try to explain the intuition that you have a climate-change-related reason not to joy-guzzle by pointing to the fact that the fully virtuous person would not joy-guzzle, but for reasons that have nothing to do with climate change.

to believe that, on average, the drive will have a negative impact on the climate. If this is true, each of us has a climate-change-related reason not to leisure drive. The fully virtuous person would of course realise this, and hence when she refrains from leisure driving, she will do so – among other things – for climate-change-related reasons. However, in this case climate change is no longer a collective impact case. The argument I am making here, as well as the inefficacy argument, only applies to genuine collective impact cases. You are free to substitute any collective impact case for the climate change example to get the gist of the argument I am making. For instance, you could consider DROPS OF WATER. In this case, the outcome will be the same whether or not you pour your pint of water into the cart. If the inefficacy argument is correct, you will have no alleviation-of-suffering-related reason to pour your pint into the cart. And, if you have no such reason, it will be of no help to consider what the fully virtuous agent would do. The fully virtuous agent would of course realise that she has no alleviation-of-suffering-related reason to pour her pint into the cart (since there is no such reason). If she were to pour her pint into the cart anyway, she would not be doing so for alleviation-of-harm-related reasons.

### **What Would a Person with Virtuous Character Traits Do?**

Should we try to understand the PRINCIPLE OF RIGHT ACTION in another way? Maybe we should take it to refer, not to what the *fully* virtuous agent would do, but rather to what an agent with virtuous character traits would do. However, if we do, we run into the same problem. According to THEORY OF VIRTUE, virtues are character traits that are generally conducive to the promotion of the good. I take it that someone with such character traits typically acts in a way that normally – that is, on most occasions – conduces to the promotion of the good.

Consider first the quite specific character trait of being a person that does not use fossil fuel powered cars unless necessary. Is this trait a virtue? Is such a character trait on most occasions conducive to promoting the avoidance of climate-change-related harms (which is a good according to our earlier assumption)? Not if the inefficacy argument is correct. If climate change and its related harms will occur whether you use a fossil fuel car or not, and if the only thing that matters is whether the occurrence of the outcome in question depends on whether or not you  $\phi$  (as the inefficacy argument assumes), having the character trait of being a person that does not use fossil fuel powered cars unless it is necessary is not on most occasions conducive to the promotion of the avoidance of climate change. It is on most occasions conducive to the promotion of actions that do not have any morally relevant impact on whether climate change and its related harms occur. So, unless we can show that having this character trait makes you perform actions that promote the avoidance of climate change and its related harms in some relevant way, the trait is not a virtue. That is, again, we run into the superfluity problem. The most

straightforward response at this point would be to show that there is some relevant causal connection between driving a fossil fuel car and climate change.

Perhaps I have misunderstood what “generally” means in *THEORY OF VIRTUE*. It is possible that refraining from unnecessary driving is generally conducive to the promotion of the good in the sense that, if everyone refrained, many climate-change-related harms would be avoided. Again, refraining from unnecessary driving could be virtuous because doing so is one of a set of acts that makes a difference for the better. If enough people refrain from leisure driving, and I am one of those people, climate change and its related harms will be less severe. The first reformulation is strongly reminiscent of the Kantian response to the inefficacy argument. I will therefore say no more about it here. Instead I refer to my comments on the Kantian response (see p. 47ff). The second reformulation reminds of appealing to what groups do together when explaining why you have an outcome-related reason to act in a certain way. I refer to my comments on that kind of solution below (see p. 62ff).

Now, the character trait we have considered so far (“being a person who does not use fossil fuel powered cars unless it is necessary”) is quite specific. Virtues and vices are typically not this specific. Rather, they are general character traits like humility, gratitude, and so on. Do we obtain a different verdict if we consider these more general character traits? We do not. For illustration, consider the virtue of loving and respecting nature (emphasised by Jamieson 2007 and Thomas E. Hill 1983), and the plausible idea that people with such a virtue would be opposed to the destruction of the natural world. Still, if it really is true that going for a drive in a fossil fuel car does not contribute to the destruction of the natural world, virtuous agents opposed to such destruction would not see a destruction-of-the-world-related reason to refrain from going for a such a drive. (Or, if they would, they would be mistaken.)<sup>10</sup> In general, if it is true that joy-guzzling makes no difference to climate change, it is unclear why refraining from joy-guzzling would be the virtuous thing to do. Similar points have been made before – for instance, by Walter Sinnott-Armstrong (2005), Joakim Sandberg (2011) and Julia Nefsky (2015, 2019). Nefsky puts it in this way:

But if my unnecessary driving or flying makes no difference with respect to climate change harms, it’s not clear why it counts as vicious, or why refraining would be virtuous.

(Nefsky 2019: 5)

---

<sup>10</sup> Hill’s (1983) original examples are different. In these, someone asphalts his garden to avoid the hassle of taking care of it, or cuts down redwood trees to make furniture. Here, it seems right to say that they destroy nature – not all of it, but a small part of it – and that a person opposed to the destruction of nature would be most likely to have acted otherwise.

The point generalises to other relevant virtues, and to other collective impact cases. For example, Nefsky (2015) considers what generous and compassionate agents would do in DROPS OF WATER and argues that it is far from clear that they would donate their pint if the inefficacy argument is correct. The problem here is that if pouring a pint truly makes no difference to the suffering of anyone, it is not generous or compassionate to do it. As Nefsky puts it:

Making a “donation” that will not do anything useful does not seem generous or compassionate; it seems foolish and wasteful.

(Nefsky 2015: 265)

In addition, as Nefsky also points out, since a virtuous person is not wasteful, we probably have to conclude that they would not add their pint to the cart.

At this point, you might think that I am aiming to prove that virtue ethics is mistaken. I am not. Rather, I want to show that turning to virtue ethics does not solve the problem posed by the inefficacy argument, but rather directs us into the superfluity problem. If there truly are no climate-change-related reasons not to go for a leisure drive, or no alleviation-of-harm-related reasons to pour a pint into the cart in DROPS OF WATER, a plausible virtue-oriented theory of ethics will not entail that there are such reasons. To paraphrase Sinnott-Armstrong (2005), changing our focus from consequences to virtues will not bring any reasons into existence. What the virtue theorist needs, and what I will try to provide later on, is an account of why we have such reasons.

Before we move on, I should mention that Nefsky understands what it is for an act not to make a difference for an outcome in another way than I do. While I take it to mean “the outcome will occur whether or not the act is performed”, she takes it to mean something like “the act does not make any difference to the outcome at all”. Thus, I agree with Nefsky that it is unclear why refraining from joy-guzzling would be the virtuous thing to do if it is true that joy-guzzling makes no difference at all to climate change. However, I also hold that joy-guzzling does make a difference of sorts for the occurrence of climate change related harms; it makes them more secure. Even if it is true that these harms would occur whether or not you joy-guzzle, they become closer to occurring (or further from not occurring) if you do.

## **Expressive Theories**

THEORY OF VIRTUE defines virtues in terms of the good, and is therefore less well suited for capturing versions of virtue ethics that do the opposite. Agent-based virtue ethics is a case in point. According to approaches of this kind, advocated by Michael Slote (2001) and Linda Zagzebski (2004), normative notions like rightness and goodness (including eudaimonia) must ultimately be explained in terms of the

virtuous motivations of agents. What matters is not whether what you do, or the character traits you have, promote the good, but rather what acting in a certain way tells us about your motives, intentions, character, and so on. This seems to be a promising solution to the inefficacy problem. For the question of whether some action of mine expresses something reproachable about me seems quite independent of the question of whether what I do makes a difference to the occurrence of some harmful outcome. If I go joy-guzzling, for instance, this might express a lack of regard for climate change and its related harms, and it might do so regardless of whether the emissions caused by this joyride make any difference as regards the occurrence of climate-change-related harms. Similarly, in DROPS OF WATER, my not donating my pint of water might express something reproachable about me regardless of whether my donation makes a difference to the suffering of anyone. Christopher Kutz (2000) makes exactly this suggestion. He writes:

In overdetermined contexts, agents can have a reason to refrain from participating in a harm, not because of the relation between this choice and an actual outcome, but because of what the choice symbolizes in their characters and commitments.

(Kutz 2000: 190)<sup>11</sup>

Sinnott-Armstrong (2005) mocks this idea (but without reference to Kutz), asking why going for a leisure drive in a gas-guzzling car would express a vice. If anything, he argues, it seems to express a desire for fun.

How can we tell whether driving a gas-guzzler for fun “expresses a vice”? On the face of it, it expresses a desire for fun. There is nothing vicious about having fun. Having fun becomes vicious only if it is harmful or risky.

(Sinnott-Armstrong 2005: 304)

Sinnott-Armstrong is correct here. Unless we can establish that there is some relevant causal connection between joy-guzzling and climate-change-related harms, it is unclear how my joy-guzzling could express something reproachable about me. I might, for instance, have thought things through carefully, not wanting to contribute to climate change and its related harms. I might have seen that there is no relevant causal connection between joy-guzzling and climate change, and decided I safely can enjoy a Sunday drive in a gas-guzzling car. Let us say I have done these

---

<sup>11</sup> This is perhaps his least well-known suggestion about how to understand reasons and accountability in collective impact cases. More famously, he suggests (concerning structured collective harm) that you are complicit in the wrong we do, or the harm we cause, if you intentionally participate in the wrong we do or harm we cause, and (concerning unstructured collective harm) that you are complicit in environmental harm if you intentionally participate in a way of life that generates such harms. More on this later (see p. 53ff).



things. If someone still reproaches me for the driving, blaming me for not caring enough about the climate, I can accurately reply that I do care about climate change. Without misrepresenting myself, I can say that if there had been any relevant causal connection between a single drive and climate change, then of course I would not have gone ahead.<sup>12</sup>

This point extends to DROPS OF WATER. If pouring a pint into the cart makes no causally relevant difference to the suffering of anyone, it is unclear why I would express a reproachable lack of concern if I did not pour my pint into the cart.

Still, it seems that I would be expressing a reproachable lack of concern for the climate if I joy-guzzled, and that I would be expressing a reproachable lack of concern for the people suffering in the desert if I kept the pint for myself. In order to explain this, however, we must presuppose that there really is a morally relevant causal connection between acting in these ways and the relevant outcomes. We must presuppose that there is such a connection between joy-guzzling and keeping my pint, on the one hand, and climate change and the people's continued suffering, respectively, on the other. More generally, when we try to explain why I have a reason to  $\varphi$  in a collective impact case by appealing to what my  $\varphi$ -ing would reveal about myself, we must presuppose a relevant causal connection between the  $\varphi$ -ing and the outcome in question. So, rather than providing a solution to the inefficacy problem, this explanation presupposes that there already is a reason. This is the superfluity problem.

There is another issue here. Even if we agree that pouring my pint into the cart is one way of showing my appropriate concern for the people suffering in the desert, it is unclear why I should show my appropriate concern by pouring my pint into the cart rather than doing something else that likewise would express my concern. As Nefsky puts it:

[The suggestion at hand] cannot explain why I have reason *specifically* to add my pint to the cart, rather than – for instance – to wear a T-shirt that says, “I support the rehydration project!” Especially insofar as we are accepting that neither amounts to doing anything instrumentally significant, wearing the T-shirt could just as well express my support, or my solidarity, as adding my pint could.

(Nefsky 2015: 263)

This is the disconnect problem. Even if we can identify reasons for action in DROPS OF WATER by appealing to the idea that we have reasons to express virtuous character traits in our actions, the reasons will not say specifically that I should pour my pint.

---

<sup>12</sup> Nefsky (2019) makes the same point.

To sum up, it seems we cannot explain the intuition that I have a reason to act in the relevant way in collective impact cases by appealing to what a virtuous agent would do, or to the idea that not acting in this way would express something reproachable about the agent's character. Have I *established* that we cannot explain the reasons intuition by appealing to virtue ethics? Not definitively. Virtue ethics is a wide and evolving field, and it may yet provide ways of explaining the reasons intuition I have not thought of. However, I suspect that any such explanation will either presuppose that there is a relevant causal connection between the relevant action and the outcome or go astray in pinpointing why we should act in this way rather than another. Differently put, I think that any successful explanation will presuppose that there is a relevant causal connection between the relevant action and the collective impact.

## Kantianism

Immanuel Kant's formula of universal law (FUL) seems tailor-made to give intuitively correct verdicts about what each of us should do in collective impact cases. It does not require me to consider the consequences of my particular act when deciding what to do. Rather, it invites me to ask: What if everybody did that? If I cannot will that everybody does what I am contemplating doing, it is impermissible for me to act in that way. For instance, since the people in the desert would continue to suffer if everybody kept their pints, and since I cannot will their continued suffering, it seems impermissible for me to keep my pint. Similarly, FUL seems to entail that I should reduce my emissions of greenhouse gases. If everyone continued using fossil fuel as usual, there would be terrible consequences. I cannot will these consequences, so it is impermissible for me to continue using fossil fuel as usual.

This sketch, however, is inaccurate. The relevant test for whether I can will that everybody acts in a certain way is not whether there will be bad consequences if they did. Rather, it is whether willing that everybody did that would result in a contradiction in the will. Perhaps surprisingly, on standard interpretations, FUL does not give the right verdict in collective impact cases. At least, not unless there is some relevant causal connection between each person's acting in the relevant way and the collective outcome.

### **The Formula of Universal Law**

FUL says that you must "Act only in accordance with that maxim through which you can at the same time will that it become a universal law" (Kant: G 4:421). There is some disagreement about how to understand this principle, but standardly it is taken to summarise a decision procedure for moral reasoning along the following

lines (see e.g. O'Neill 1989, Rawls 1989, and others): First, formulate a maxim that describes your reason for acting. Second, consider a world where this maxim is made universal law – that is, a world where everybody always acts on this maxim when they are in the relevant circumstances. Third, ask whether this world is conceivable. If it is inconceivable, there is a perfect duty not to act on the maxim. Fourth, if the world is conceivable, ask whether you could rationally will such a world. If not, there is an imperfect duty not to act on the maxim. If there is neither a perfect duty nor an imperfect duty not to act the maxim, it is permissible to act on it.

Perfect duties are strict and exceptionless. They require that you never act on the relevant maxim in the relevant circumstances. For illustration, there is a perfect duty not to make lying promises. In a world where everyone acts on the maxim “Make a lying promise” whenever convenient, no one would believe in promises, and so it would be impossible to even make a promise. A world where only convenient promises are kept is inconceivable.

As many have pointed out before me, there is typically not a perfect duty to act in the relevant way in collective impact cases.<sup>13</sup> For instance, I do not have a perfect duty to refrain from driving a fossil fuel powered car, or to pour my pint into the cart in DROPS OF WATER. Why? Because a world where everyone drives fossil fuel powered cars when they are in the relevant circumstances is conceivable, as is a world where everyone declined to pour their pints into the carts when they find themselves in circumstances similar to those in DROPS OF WATER. These are morally deplorable worlds, but they are not inconceivable.

How about imperfect duties? While perfect duties are strict, we have some leeway when it comes to imperfect duties, which require us to sometimes, and to some extent, to avoid acting on the relevant maxim when we are in the relevant circumstances.<sup>14</sup> Kant considers the example of benevolence. It is possible to imagine a world where everyone acts on the maxim “Do not give help to others”. That is a conceivable world, so there is no perfect duty to refrain from acting on that maxim. However, Kant argues, in another sense you cannot will that everyone acts on this maxim, for if they did, you would never yourself receive help when you needed it (Kant: G 4:423). A world where no one ever gives help to others would be a world where important aims of yours would be frustrated. So, even though I am permitted to sometimes act on the maxim “Do not give aid to others”, I am not permitted to always act on this maxim. Sometimes, and to some extent, I should act benevolently.

---

<sup>13</sup> See e.g. Kutz (2000: 133-35), Sinnott-Armstrong (2005), Sandberg (2011) and Nefsky (2015).

<sup>14</sup> See e.g. Kant (2002/1785). Kant writes that a perfect duty “permits no exception to the advantage of inclination” (G 4:421n), leaving the reader to infer that imperfect duties do allow some such exception.

Do you have an imperfect duty to act in the relevant way in collective impact cases? Some believe you do. Thus Baylor Johnson (2011) and Christian Baatz (2014) argue that you have a Kantian imperfect duty to reduce your emissions of greenhouse gases. According to Johnson, you have an imperfect rather than a perfect duty to reduce your emissions, since “the burden on individuals doing this unilaterally is too great given the odds against their sacrifice achieving much in the absence of collective schemes” (2011: 151).<sup>15</sup> Baatz (2014) agrees. The argumentation here is quite un-Kantian. Typically, at least, the range of perfect duties you have is not affected by whether they place great burdens on you or not. Rather, what matters is whether a world where the relevant maxim is universally followed is conceivable. Further, while Johnson in a typical Kantian vein thinks that “the amount of reduction cannot be specified and is left to the judgment of the individual” (2011: 151), Baatz makes the somewhat less Kantian argument that you have an imperfect duty to reduce your emissions “as far as can reasonably be demanded” (2014: 10). When it is spelled out in this way, the duty seems to require more than is normal for an imperfect duty. These details are of limited importance given that Johnson and Baatz could still be right that there is a Kantian *imperfect* duty to reduce our emissions of greenhouse gases. This is the possibility I will consider in what follows.

FUL does not necessarily entail that you have an imperfect duty to reduce your emissions of greenhouse gases. Judging by his comments on the imperfect duty of benevolence, Kant’s idea is that I cannot will a maxim to be universally followed if some relevant aim of *mine* would be frustrated in a world where everyone acts in accordance with this maxim.<sup>16</sup> However, my aims would not be frustrated in a world where no one refrains from reducing their emissions of greenhouse gases (here I am setting aside my altruistic aim that people should not suffer climate-change-related harms). Most climate-change-related harms, at least those severe enough to threaten to frustrate any important aim I have, are likely to occur long after my death. So, acting on this maxim would not generate a contradiction in my will. Similarly, as Nefsky (2015) argues, no aim of mine would be frustrated if everyone kept their pints in DROPS OF WATER. I therefore have no imperfect duty not to keep my pint in this case.<sup>17</sup>

---

<sup>15</sup> Johnson had earlier argued for a slightly different position (see Johnson 2003).

<sup>16</sup> Famously, this lead Schopenhauer (1969/1818-19) to reject the categorical imperative. Schopenhauer thought that compassion was the ground for all morality, and that morality therefore could not be derived from egoism.

I am not sure whether Kant would say that I cannot will a maxim to be universally followed if *any* aim of mine would be frustrated in a world where everyone acts according to this maxim, or if he would say that it has to be *some important* aim of mine, or if he would have some other way of distinguishing the relevant aims. What I say in the main text cover all these possibilities.

<sup>17</sup> Parfit (2011: 334-38) makes a similar point, calling it “The Non-Reversibility Objection”.

Were the maxims we considered too narrow? As O'Neill (1989), Rawls (1989), Wood (1999), Parfit (2011) and others have argued, we will find countless counterexamples to FUL if we apply it to excessively specific maxims.<sup>18</sup> Maybe the relevant maxims are not "Do not reduce your emissions" and "Keep your pint", but some more general maxims like "Contribute to the harming of others" whenever convenient, or "Do not give aid to others" whenever you are in the relevant circumstances. There does seem to be an imperfect duty not to act on these maxims. Many of my important aims will be frustrated in a world where everyone contributes to the harming of others whenever it is convenient, and the same goes for a world where no one gives help to others.<sup>19</sup>

Now, just when we have found maxims that are general enough to entail imperfect duties, we run into another problem. For it seems that these maxims do not apply to the examples under consideration. If there is no causal connection between going for a single drive and climate-change-related harms, why would going for such a drive count as contributing to the harming of others? If we agree that making a difference to the occurrence of an outcome is the only causal connection there is, it seems we must conclude that a single drive would not contribute to climate change and its related harms, and in turn that the duty not to contribute to the harming of others is irrelevant to the question whether it is permissible for me to go for such a ride.<sup>20</sup> (I am here assuming that contributing is a causal notion). The same point can be made about DROPS OF WATER. Unless there is some causal connection between pouring a pint into the cart and the alleviation of harm, pouring a pint into the cart would not count as helping anyone. So, again, FUL falls short of implying that I have an imperfect duty to pour my pint into the cart.

It seems that we have to choose between, on one hand, maxims like "Do not reduce your emissions" and "Keep your pint" that I can will everyone to follow (and therefore do not entail duties of any sort) and, on the other hand, maxims like "Contribute to harm to others" and "Do not give help to others" that I cannot will everyone to follow (and therefore do entail duties) but also fail to apply to collective impact cases unless we can show that there is a causal connection between acting in the relevant way and the collective outcome. The obvious question at this point is: Are there any other maxims that are general enough to entail that I cannot will their universalisation but do not incorporate any causal notion such as "contribute" or "benefit"?

---

<sup>18</sup> See in particular O'Neill (1989: 83-86), Rawls (1989: 86), Wood (1999: 102-07), Parfit (2011: 289-300). Wood argues that maxims should neither be too specific nor too general.

<sup>19</sup> Still, there is no perfect duty associated with these maxims. A world where everyone follows the maxims when in the relevant circumstances is conceivable, if also awful.

<sup>20</sup> I should remind you that I take making a difference to the occurrence of the outcome to mean that it is true that the outcome had not occurred if the action had not been performed. (See Chapter 1.)

Perhaps such maxims can be identified in all collective impact cases. If that were so, we could show that FUL entails that I have a duty not to act in the relevant way in all collective harm cases without also showing that there is a causal connection between acting in the relevant way and the collective outcome. Although I cannot think of any such maxim, I cannot rule out the possibility that they exist. Still, the arguments that I have presented here, while they are not conclusive, point in the direction of the following result: Unless we can show that there is a causal connection between acting in the relevant way and the collective outcome in collective impact cases, FUL will not entail that there is an imperfect duty to act (or refrain from acting) in this way.

### **A Revised Version of FUL**

We might tweak FUL so that it says I have imperfect duties not to act on a certain maxim whenever universal compliance with this maxim would have bad consequences of the right sort. Perhaps general compliance with a particular maxim would undermine the possibility of rational agency. Perhaps if everyone acted on the maxim, human dignity would be compromised, or there would simply be bad enough consequences. We could then argue that there is an imperfect duty not to act on the maxim “Keep your pint” in DROPS OF WATER, on the basis that if everyone involved in the situation did that, there would be bad consequences. Human dignity would be compromised, for instance. This revised version of FUL echoes the version I described in the introduction to this section: If I cannot will that everybody acts in a certain way because of the bad consequences this would lead to, I have a duty not to act in that way. It also reminds us of what Glover (1975) calls “the generalisation test”, and Kutz (2000) also understands Kantian imperfect duties along these lines.<sup>21</sup>

---

<sup>21</sup> Parfit (1984) argues that a Kantian agent would not necessarily be motivated to, say, pour his pint of water into the cart in DROPS OF WATER, because he might think the following: “If my contribution would make no difference, I can rationally will that everyone else does what I do” (1984: 67). Parfit’s idea is not that the Kantian agent might think it permissible to keep his pint because if everyone kept their pints no aim of his would be frustrated. Rather, it is that the Kantian agent might think it permissible to keep his pint because a world where everyone keeps their pints is still a world where no one by himself makes a difference to the suffering of anyone. This agent appears to be less than fully Kantian. FUL requires us to consider what would happen if everyone kept their pints, and if everyone kept their pints the suffering would continue. A fully signed-up Kantian (who believes in the revised version of FUL) would take this to entail an imperfect duty not to keep his pint.

Kutz (2000) argues that what I have called the revised version of FUL does not entail that I have an imperfect duty to act in the relevant way in collective impact cases. His arguments miss their target for the same reason as Parfit’s argument does. In a world where everyone in the relevant circumstances performs the relevant action, the consequences will be dire, and therefore FUL (on the current interpretation) will condemn acting in this way.

The revised version of FUL gives the intuitively correct verdict in the climate change case. Sometimes, and to some extent, I should reduce my emissions of greenhouse gases. We can argue on this basis, for instance, that driving someone to an emergency of some sort is permitted, but that driving with less serious aims is not. The revised version of FUL also gives the roughly correct verdict that I have an imperfect duty to pour my pint into the cart in DROPS OF WATER. (This verdict is only roughly correct, since it seems that I should always pour my pint into the cart in a situation like DROPS OF WATER, not only do so sometimes and to some extent.)

Unfortunately, however, this version of Kantianism gives intuitively unappealing verdicts in cases of overdetermination. Consider a variant of DROPS OF WATER where we are 15,000 pint holders, where the cart still cannot take more than 10,000 pints, and where 10,000 pint holders have already donated their pints. Here, it seems I do not have a duty to pour my pint into the cart. Doing so would simply be a waste of resources. Still, the tweaked version of FUL entails that I have an imperfect duty to pour my pint of water into the cart. Since there would be bad enough consequences in a world where everyone kept their pints (the people in the desert would continue to suffer, human dignity would be eroded, and so on), I have an imperfect duty to donate my pint. That is, sometimes, when I find myself in a case like DROPS OF WATER, I should pour my pint into the cart even if it is full. This verdict seems mistaken.<sup>22</sup>

There is, however, a way out of this problem. We could argue that the relevant maxim in this case is not “Keep your pint”, but rather something like “Do not give aid to others”. If we do this, we can explain why I do not have an imperfect duty to pour my pint into the cart when the cart is already full. In these circumstances, pouring a pint into the cart would not aid anyone. While pouring a pint into the cart before it is full might count as helping to alleviate suffering, making the one pint’s worth of water overflow is a completely superfluous thing to do. Still, we will only be in a position to appeal to this explanation if we can show that pouring a pint of water into a cart that is not yet full contributes to the alleviation of harm while pouring it into a cart that is full does not. Again, we see that the Kantian response to the inefficacy argument works only if there is a relevant causal connection between acting in the relevant way and the outcome in question.

---

<sup>22</sup> There are also other reasons to be sceptical about the generalisation test. Sandberg (2011) rejects it, arguing that it is unreliable. As he puts it: “It would certainly be disastrous if everyone were celibate because no future generations would then be born. Similarly, it would have devastating effects if everyone lived in Sweden, since even though Sweden is rich in natural resources, these resources in no way could sustain the Earth’s entire population. Yet it seems absurd to say that it is morally *wrong* to be celibate or to live in Sweden.” (238). Others have presented similar arguments (e.g. Lyons 1965).

## The Formula of Humanity

Consider instead Kant's second formulation of the categorical imperative: the formula of humanity. This says that you should "Act so that you use humanity, as much in your own person as in the person of every other, always at the same time as end and never merely as means" (Kant: G 4:429). Could this formula explain why you have a reason to refrain from driving a fossil fuel powered car? It seems not. If there really is no causal connection between going for a single drive and climate change, it is unclear, at best, why such a drive would count as treating those who suffer climate-change-related harms as mere means. The same reasoning applies to DROPS OF WATER. Unless we can show that there is some relevant causal connection between pouring a pint into the cart and alleviation of the people's harm, it seems that you are not treating them in any way at all by pouring your pint into the cart. *A fortiori* you are not treating them as mere means. As Nefsky (2015) puts it: "Unless your act would play some significant role in causing the outcome, it does not seem we can appeal to the Formula of Humanity to explain why you ought not to do it" (266).

This pattern seems to repeat itself when we consider influential interpretations of Kant's formula of humanity. O'Neill (1985) and Korsgaard (1996) argue, for instance, that ultimately it is coercion and deception that make it impermissible to act on certain maxim. However, going for a leisure drive in a car powered by a fossil fuel is neither deceptive nor coercive. You might think that you contribute to the coercing of future generations into living a worse life by going for such a drive. Maybe you do. However, if there is no causal connection between a single drive and climate-change-related harms, it seems hard to establish that going for such a drive coerces future generations in any way at all. Similarly, if there really is no causal connection between pouring a pint into the cart and the alleviation of harm caused by thirst, you do not seem to be coercing the people suffering from thirst to continue doing so by keeping your pint.<sup>23</sup>

## Complicity

In some cases, it seems that you are morally responsible for a harm in virtue of participating in a collective action leading to that harm even though what you did made no difference to the occurrence of the outcome. Consider, for instance, the following case introduced by Kutz (2000: 116-24), here presented in shortened form:

---

<sup>23</sup> I have not here considered the third and fourth formulations of Kant's categorical imperative: the autonomy formula and the kingdom of ends formula. That will have to be the topic for another day.



THE DRESDEN BOMBINGS was an Allied aerial bombing attack on the city of Dresden in February 1945. The attack, which involved a series of mainly incendiary bombing raids, has been described as inhumane and strategically useless. The bombing and the resulting firestorm killed approximately 35,000 civilians and devastated the city centre. There were more than a thousand planes involved and eight thousand crewmembers directly participated as pilots, gunners, navigators and bombers.

Here, the firestorm, and the ensuing death and destruction, would have occurred even if any particular crewmember had refused to participate, or had dropped their bombs in a deserted area outside the city. So, as Kutz argues, if all that matters for accountability is whether the outcome had been different if you had acted in a different way, “each bomber can truly reply to the victims or their survivors, ‘Why blame me? I have not caused your suffering, nor made you worse off’” (122). Similarly, if all that matters for accountability is whether the outcome had been different if you had acted in a different way, the crewmembers do not have any grounds for regret. Still, it seems that each crewmember is accountable for the harmful outcome. Several crewmembers also saw themselves as such. Some had serious regrets over what they done. Others were pleased about the destruction they had caused to the hated enemy. So, the idea that all that matters for accountability is whether the outcome had been different if you had acted in a different way seems mistaken.

On Kutz view, you might be accountable for the outcome of a collective action if what you did relates to the collective action in a relevant way, and that this is true even if what you did made no difference to the outcome. He suggests the following principle to describe the relevant relation:

THE COMPLICITY PRINCIPLE: (Basis) I am accountable for what others do when I intentionally participate in the wrong they do or harm they cause. (Object) I am accountable for the harm or wrong we do together, independently of the actual difference I make.

(Kutz 2000: 122)<sup>24</sup>

This principle gives the right verdict about THE DRESDEN BOMBINGS. Each crewmember intentionally participated in the collective action that brought about

---

<sup>24</sup> THE COMPLICITY PRINCIPLE is closely tied to Kutz’ account of collective (or joint) action. On this account, you participate in a collective action when you intentionally participate in the relevant act. As Kutz puts it, “A set of individuals jointly G when the members of that set intentionally contribute to G’s occurrence by doing their particular parts, and their conceptions of G sufficiently and actually overlap” (Kutz 2000: 103). Kutz makes clear that he does not understand what it is to contribute to G’s occurrence in a causal way. What’s doing the work here is that the members of the set have overlapping collective intentions on which they act.

the firestorm and resulting devastation, and so each is accountable for this horrific outcome.<sup>25</sup>

THE COMPLICITY PRINCIPLE is not about reasons, but accountability. However, it is easy to see how we could build on it to explain the reasons intuition. We could argue that I have a reason not to intentionally participate in the wrong others do, or harm they cause, since if I do, I will be accountable for this wrong or harm. If this argument is sound, I have a reason to refrain from participating in harmful collective actions even if my particular act makes no difference to whether or not the harmful outcome occurs. This seems to be a promising way to explain the reasons intuition.

The more general point is this: I might have a reason not to partake in a collective action that brings about harm even if my act makes no difference to whether or not the harm occurs. You do not need to subscribe to Kutz' complicity principle (or his account of collective action) to accept this idea. You can press into service whatever account of collective action (or complicity) you prefer.<sup>26</sup>

Problematically, this line of reasoning cannot explain the reasons intuition in all collective harm cases. For some collective harms are not the intended result of a collective action. HARMLESS TORTURERS is a case in point. So is climate change. In HARMLESS TORTURERS, the victim's pain is not brought about as a result of a collective action.<sup>27</sup> Climate change, in turn, is not the result of one large, coordinated action, but rather the result of many different agents doing ordinary things. These are *unstructured collective harms*.<sup>28</sup> Kutz recognises this problem. The remainder of this section considers his and Brian Lawson's (2013) attempts to extend THE COMPLICITY PRINCIPLE to cover such unstructured harms.

---

<sup>25</sup> It might be objected that it is excessive to hold each crewmember fully accountable for this outcome. For instance, it might be insisted that no crewmember is accountable to the same degree as those who planned this atrocity. Kutz would agree. THE COMPLICITY PRINCIPLE does not say anything about the nature or degree of the accountability. It does not imply that each crewmember is as accountable for the devastation as the person, or people, who planned the collective action.

As I understand Kutz, to be accountable for an outcome, or action, is to be morally responsible for it. When you are accountable for an action or outcome others are warranted in reacting positively or negatively to you in virtue of this action or outcome. The exact nature and degree of the warranted reactions depend on what you are accountable for (the object), the relation in virtue of which you are accountable (the basis), and the relation between you and the reacting agent. A surviving victim may, for instance, be warranted in reacting more fiercely to a particular crewmember than they do to an onlooker or the crewmember's superior. (See, in particular, Kutz 2000: ch. 2.)

<sup>26</sup> You may prefer to think of collective action along the lines of Bratman (1993, 2014), Gilbert (1989, 2014), Searle (1990), Ludwig (2016), French (1984) or List and Pettit (2011). Some of these accounts lead to additional difficulties with the suggestion at hand.

<sup>27</sup> This example was introduced on p. 23.

<sup>28</sup> There can of course also be unstructured collective benefits. DROPS OF WATER illustrates this.

## Kutz on Unstructured Collective Harms

Kutz considers the following case:

[CFC COOLANTS:] Automobile air conditioners are a significant factor in the release of CFCs, and so are a prime contributor to the widening of ozone holes. Fortunately, CFC-free coolants are available, but at a much higher cost and with substantially less cooling power.

Let us say, for the purposes of the example, that American drivers of CFC-cooled cars contribute 25% of the CFCs released into the atmosphere, and that the increased CFC emissions globally have been linked to 4,000 additional skin cancer cases in Northern Australia because of a hole in the ozone layer. And so, let us say 1,000 skin cancers can be causally attributed to American drivers.

(Kutz 2000: 171)

Kutz argues that the group causing the harm is clearly identifiable from the victim's perspective: it is the polluters. They are "engaging in a concrete way of life that generates these harms" (186). Still, he continues, the individual drivers will not always see themselves as connected to harm. They may not think about the harm at all, and if they do, they might still think their use of air conditioning is inconsequential as regards the harm. If they do, Kutz says, the harm is a "product of a kind of ethical *anomie*, of a tendency to regard one's moral relations to the world as essentially isolable" (187).

The challenge, as Kutz sees it, is to make these drivers see themselves as participants in a shared activity – a shared activity that inadvertently causes harm. That is, we must make them see things roughly as the victims do. If we manage to do this, we might elicit in them a sense of accountability, and this sense of accountability might in turn motivate them to change their habits, and to accept political solutions aimed at redressing the harm. The key to doing this is to make people realise that they already "share an objectively determinate and highly interdependent way of life" (188). Their lifestyle is only possible, for instance, given relatively cheap fuel and "disguised public subsidies of automobile travel" (188), and these social conditions are in turn produced by those citizens who enjoy driving and their allied organisations and corporate agents. In other words, their way of life is enabled by certain socioeconomic structures which, in turn, are sustained and reproduced by those who favour this way of life.<sup>29</sup> If people saw this, they might acquire the requisite sense of accountability.

---

<sup>29</sup> As Hormio (2017) argues, there seems to be substantial overlap between Kutz's proposal and Young's social connection model. I discuss the social connection model in Gunnemyr (2020).

Kutz's strategy might work in combatting some environmental harms, but it does not apply successfully to all collective impact cases. The strategy crucially relies on there being a pre-existing way of life that causes harm. However, in some collective impact cases, there is no pre-existing way of life of this sort. DROPS OF WATER illustrates this. Nothing in the example indicates that we share a way of life. We can readily imagine that we – the pint holders – find ourselves in the unlikely circumstances described in the example out of sheer, bizarre coincidence. Obviously, if we do not share the required pre-existing way of life, we cannot realise that we do.

Moreover, even if we do share a way of life without thinking about it – we might all come from the same country and be living highly interdependent lives – it is unclear why realising this should be our top priority in this one-shot case. What we should do is simply pour our pints into the cart. If our way of life were the reason why the people in the desert are suffering, we might have an additional reason to do something about their suffering. But let us assume that this is not the case. Our way of life has nothing to do with their suffering. Given this, it is unclear, at best, why we should start seeing ourselves as sharing a determinate and highly interdependent way of life, enabled by certain propitious socioeconomic structures. Again, what we should do is simply pour our pints into the cart.

Kutz is nonetheless correct in saying that the “tendency to regard one’s moral relations to the world as essentially isolable” is problematic. Suffering from such ethical anomie, each of us might think along the following lines: “People’s suffering will be just the same whether or not I pour my pint into the cart, so why should I bother doing so?”<sup>30</sup> If we see matters along these lines, we will probably not think of ourselves as having a reason to pour our pints into the cart. Likewise, we will be unlikely to think of ourselves as accountable for the continued suffering of the people in the desert if we fail to pour our pints into the cart. In “You Just Didn’t Care Enough” (Chapter 11), Touborg and I argue that it is mistaken to think along these lines.

Kutz’s solution also fails to identify those accountable for the harm in CFC COOLANTS. He argues that everyone who shares the relevant way of life should foster a sense of accountability for harm. However, not everyone who shares this way of life will necessarily use CFC coolants. Some may not have an air conditioner, and some may have chosen an environmentally friendly alternative. We might well ask why these people should foster a sense of accountability for ozone layer depletion and its related harms. It seems more accurate to say that those who are using CFC coolants (or selling or producing them) should foster that sense.

---

<sup>30</sup> At least, we might think so if one extra drop of water has no chance of making a difference to anyone’s suffering of thirst. I discuss this issue in Chapters 7 through 9.

David T. Schwartz (2017) makes the same point in relation to consumer choices. Considering whether we can hold consumers accountable for harms like sweatshop labour and environmental degradation on the grounds that they intentionally participate in a consumer culture which in turn causes these harms, he writes:

The broad appeal to consumer culture would seem to imply that even consumers who attempt to buy ethical products (e.g., green products, fair-trade products) are just as culpable as everyone else simply because they, too, are participants in the consumer culture. The issue is not consumption per se but specific purchases and types of purchases.

(Schwartz 2017: 101)

There is also the reverse problem. You might contribute to unstructured collective harm without participating in a harmful way of life. Suppose you have left society behind and are now living on your own in a forest, neither depending on nor reproducing any socioeconomic structures. Here, you find an abandoned coal mine, and every night you burn as much coal as you can just for the fun of it. Here, it seems that you have a climate-change-related reason not to burn coal. However, it cannot reasonably be argued that you should realise that you are participating in a harmful way of life. You do not participate in any way of life, at least in the sense implying that you depend on and reproduce certain socioeconomic structures.<sup>31</sup>

The issues I have raised here are all versions of the disconnect problem. In focusing on participation in harmful ways of life, Kutz fails to reliably distinguish, in some collective impact cases, those who are accountable for harm from those who are not.

As a final point, Kutz also suggests that there might be *symbolic accountability* (2000: 190-91). Others might be warranted in reproaching me if I participate in a harmful way of life since my doing so might express something reproachable about my character. For instance, by joy-guzzling, I might express a reproachable attitude towards the climate. There is something to this. Still, this cannot explain accountability in all collective impact cases. In some such cases, as we have seen, there is no harmful way of life in which I participate. For further comments, I refer the reader to my discussion of expressive theories earlier in this chapter.

---

<sup>31</sup> Banks (2010) raises an additional problem. She questions why our accountability practices would change when considering unstructured collective impact cases. It seems that there should be some common ground on which we could hold individuals accountable in all collective impact cases. I agree, but I set this objection to the side here. Hormio (2017) considers and answers other objections to Kutz's treatment of unstructured collective harm cases, but she does not satisfyingly answer the kind of worries I have raised here.

## Lawson on Unstructured Collective Harms

Lawson (2013) argues that we should revise THE COMPLICITY PRINCIPLE to say the following:

MODIFIED COMPLICITY PRINCIPLE: (Basis) I am accountable for what others do when I knowingly contribute to a harmful outcome that results from our collective contributions. (Object) I am accountable for the harm or wrong we do together, independently of the actual difference I make.

(Lawson 2013: 234)

On this proposal, there need not be any collective endeavour, or way of life, in which I intentionally participate. It suffices that I knowingly contribute to a harmful outcome in order to be accountable for it. This proposal gives intuitively correct verdicts about accountability in the cases we have discussed. When you burn your coal, you knowingly contribute to climate change and its related harms. You are therefore accountable for the harms on the MODIFIED COMPLICITY PRINCIPLE. Similarly, the drivers who use CFC coolants are accountable for ozone layer depletion and harms related to it since they contribute to these harms – at least, they are accountable insofar as they do this knowingly. By contrast, drivers who use environmentally friendly coolants, or do not use air conditioners at all, are not accountable for harm since they do not contribute to them. And so on.

However, I fail to see how you can knowingly contribute to a harmful outcome without actually contributing to it. In order to knowingly contribute to some outcome, there has to be some causal connection between what you do and this outcome. There has to be some relevant sense in which, for instance, using CFC coolants in your car contributes to ozone layer depletion and its related harms. In the absence of such a connection, the MODIFIED COMPLICITY PRINCIPLE will not work. That is, this principle faces the superfluity problem.<sup>32</sup>

Neither Kutz nor Lawson accepts the idea that a single driver using CFC coolants is causing ozone layer depletion and its related harms. Kutz concedes that there might be a causal connection of sorts here, but not that there is any morally relevant causal connection. I consider Kutz's arguments for the claim that there is no morally relevant causal connection in cases of unstructured collective harms more closely in the next chapter, and argue that they are mistaken.

---

<sup>32</sup> For further discussion of Lawson's proposal, see Petersson (2013).

## Reasons to Take Collective Action

Agents might have reasons to act together to combat unstructured collective harms. They might have reasons to take collective action in order to prevent further climate change, and they might have reasons to decide upon a plan to alleviate the thirst of those suffering in the desert in DROPS OF WATER. This, roughly speaking, is Tracy Isaac's (2011) preferred solution to the problem of unstructured collective harm. Stephanie Collins (2019) and Frank Hindriks (2019) similarly argue that individuals sometimes have duties to collectivise or coordinate, and Virginia Held (1970) and Larry May (1990) put the parallel argument that individuals might be morally responsible for not acting collectively. Could such reasons explain the reasons intuition?

I agree that people might have reasons to take collective action, but I disagree that these reasons explain the reasons intuition. I have two basic concerns. First, the reasons intuition is neither necessarily nor primarily about reasons to take collective action. It is, for instance, not about having reasons to take collective action in order to prevent further climate change. It is about having reasons to reduce one's own carbon footprint. Similarly, the reasons intuition does not say that you have a reason to cooperate with others in DROPS OF WATER, but that you have a reason to donate your pint.

In many collective impact cases, you both have an outcome-related reason to act collectively and a reason to  $\phi$ . Take climate change, for instance. Here, it seems that you both have a reason to minimise your use of fossil fuels *and* a reason to act together with others in order to address climate change. These reasons are not mutually exclusive.

In some cases, you might have one of the reasons, but not the other. If you live in a repressive society where no form of cooperation or organisation that is not state-sanctioned is forbidden, you might lack a reason to act together with others to combat climate change. In such a society, there might be no possibility that you would make a difference to climate change by taking collective action. If you tried, you would simply end up in prison. Even so, you might have a climate-change-related reason to minimise your own use of fossil fuels. Again, in a science fiction scenario, a neuroscientist might have implanted a chip in your brain, making it impossible for you to cooperate with others. Here, it seems, you lack a reason to cooperate with others, since doing so is simply impossible. Still, it seems that you have a climate-change-related reason to reduce your own usage of fossil fuels.

Conversely, you might have an outcome-related reason to cooperate with others without having a reason to perform some other specific act. Consider, for instance, Held's (1970) case, in which a man can be saved if only some people cooperate, which I have abridged:

REMOVING BEAMS: A man is trapped inside a collapsed building. He needs immediate assistance: unless various beams are removed, he will slowly bleed to death. There are three unacquainted pedestrians at the scene. They could save the man by removing the beams, but the removal of any one beam would require the strength of all three. Regrettably, the three pedestrians do not agree on how to proceed. One argues for first removing beam 1, another for first removing beam 2 and a third argues for first removing beam 3. While they argue, the man slowly bleeds to death.

Here, no pedestrian had a saving-of-the-man-related reason to remove any particular beam. They did, however, have a reason to cooperate, either by deciding on a decision-procedure (as Held argues), or by simply acting more responsively with one another. When one person starts lifting one beam, the others might have responded by helping with that beam instead of arguing that they should proceed in some other way. Since reasons to  $\varphi$  and reasons to take collective action can come apart, you cannot explain one by appealing to the other.

My second concern is that appealing to reasons to take collective action does not always help us to avoid the inefficacy problem. You might think that I have reasons to start, or join, a local environmental activist group, working together to combat climate change. But what we do in this small group will not make a difference to any climate-change-related harm, so if the inefficacy argument is correct, I will lack a reason to start or join such a group. Again, you might think that I should work to transform some existing larger political body in order to mitigate climate change. For instance, you may think I should try to influence Sweden's climate policies by protesting, or petitioning, or joining some political group. That is, you might agree with Sinnott-Armstrong that it is "better to enjoy your Sunday drive while working to change the law so as to make it illegal for you to enjoy your Sunday driving" (2005: 312), or with Young (2011) that I have reasons to work with others to ameliorate the harmful social structures in which I participate.

However, if the inefficacy argument is correct, I will probably lack such reasons. My protesting will not make any difference to whether Sweden changes its climate policies. It is true that if enough people were to protest, Sweden would be likely to change its policies, but if this were to happen it would do so whether or not I participate in the protests. My petition might be approved, but only if there is a majority in the government that approves of it. And, if there is a majority in the government that wants more progressive climate policies, they will implement such policies whether I file my petition or not. Doubtless, if enough people organise and act together for a change, Sweden will probably change its policies, but if this happens it will happen whether or not I participate in this movement.

The inefficacy problem resurfaces on many levels. For instance, even if you think that there is some prospect that my protesting, petitioning or joining or forming a political movement could affect Sweden's climate policies, it seems that this would be completely inefficacious. On a global scale, Sweden's emissions are a drop in



the ocean. Even the most progressive climate policies and regulations in Sweden would make little difference to climate change.

Now, it does seem that you do have climate-change-related reasons to act together with others to challenge climate change, and that you have reasons to vote for a party that wants to implement progressive climate policies. However, in order to show that you have such reasons, we first have to show where the inefficacy argument goes wrong.

## Membership in a Group

So far, I have considered a range of proposals as to why you have a reason to act in a certain way in non-threshold cases like DROPS OF WATER and HARMLESS TORTURERS, but I have not yet considered Parfit's proposals. In essence, Parfit claims that we must *either* hold that there can be imperceptible harms and benefits *or* accept that an action can be wrong because it is one of a set of acts that together harm people.<sup>33</sup> He adopts the first option. He says that you have a reason not to flip a switch in HARMLESS TORTURERS because your doing so would harm the victim imperceptibly, and that you have a reason to pour your pint into the cart in DROPS OF WATER because your doing so would benefit the people in the desert imperceptibly.<sup>34</sup> However, he also recognises that the claim that there are imperceptible harms and benefits is controversial, and that it therefore might be better to appeal to what groups together do.<sup>35</sup> This section discusses the appeal to what groups together do. Parfit spells out this idea in the following way:<sup>36</sup>

---

<sup>33</sup> I have here simplified things a little. Parfit (1984) says that we must either reject (A) or (B), where (B) is the claim that the at-least-as-bad-as-relation and the not-worse-than-relation are transitive. If we reject (B), he argues, we must accept that an act might be wrong because it belongs to a set of acts that together harm people (i.e. the claim I state). The reason we must reject either (A) or (B) is that otherwise we must conclude the victim in HARMLESS TORTURERS is no worse off when 1000 switches are flipped than he is when 100 switches are flipped, which is absurd. See Parfit (1984: 78-82).

<sup>34</sup> For instance, he writes that we "must reject either (A) or (B). Which should go? I reject (A)" (1984: 79). He also states that he "prefer[s...] to appeal to the effects of single acts" (1984: 82). See also Parfit (1986: 847).

<sup>35</sup> Parfit (1984: 82).

<sup>36</sup> In *Reasons and Persons*, Parfit does not appeal to (C7), but proposes another principle, (C12) (and the similar principles (C10) and (C11)), to capture our intuitions about what we have reasons to do in cases like DROPS OF WATER and HARMLESS TORTURERS, given that we do not agree that there are imperceptible harms and benefit. However, in his comments on Gruzalski's (1986) criticisms of *Reasons and Persons*, he later says that he should instead have appealed to (C7). See also Parfit (1986).

(C7) Even if an act harms no one, this act may be wrong because it is one of a *set* of acts that *together* harm other people. Similarly, even if some act benefits no one, it can be what someone ought to do, because it is one of a set of acts that together benefit other people.

(Parfit 1984: 70)

We might ask what it means for an act, or set of acts, to harm or benefit someone. Here, Parfit appeals to the following principle:

(C6) An act benefits someone if its consequence is that someone is benefited more. An act harms someone if its consequence is that someone is harmed more. [...]

(Parfit 1984: 69)

I take it that this principle applies to sets of acts as well as acts.<sup>37</sup> I also take it that it is implicit in (C6) that the relevant comparison is with what would have happened if the act or act set had not occurred. That is, we are to compare what happens in the actual world where the act or act set occur with the closest counterfactual world where it or they do not. If someone is harmed more in the actual world than in that counterfactual world, the act or act set harms this person. Likewise, if someone is benefitted more in the actual world than in that counterfactual world, the act or act set benefits this person. For short, what matters in (C6) is whether the outcome would have occurred if the act or act set had not.

Anne Schwenkenbecher (2014) similarly appeals to what groups do together. She argues that as individuals we have reasons to reduce our individual emissions of greenhouse gases since “our individual actions are potentially harmful to others not by themselves, but because they are part of a set of similar actions which together cause harm” (176). Unlike Parfit, she spells out her argument in terms of *causing harm* to others rather than *harming* others. However, insofar as you take causation to consist in counterfactual difference-making, this makes no difference. In both cases, what matters is whether someone is harmed more in actual world than in the closest counterfactual world where the relevant set of acts does not occur.<sup>38</sup>

---

<sup>37</sup> What Parfit says in his “Comments” (Parfit 1986) would not have made sense otherwise.

<sup>38</sup> Like Parfit (1984), Schwenkenbecher appeals to (C10). Here, I concentrate instead on (C7). I explain why in footnote 36, this chapter. All of the points I make here could also be made about (C10) (and (C11) and (C12)).

Besides appealing to (C10), Schwenkenbecher argues that we might have a reason not to reduce our carbon footprint because our not achieving the reduction might actually spur others to reduce their emissions, and perhaps initiate collective action aimed at tackling climate change. I will discuss indirect consequences later in this chapter.

I have two basic concerns about the idea that an act might be wrong because it belongs to a set of acts that together harm people, and the similar idea that an act might be what you ought to do because it belongs to a set of acts that together benefit people. First, how do we establish whether any given act belongs to the relevant set? Second, how do we establish that this set is harmful (or beneficial)? I will address these questions in turn.

### **Does My Act Belong to the Set?**

On an intuitive level, it seems obvious that flipping a switch in HARMLESS TORTURERS belongs to a set of actions that harms the victim, and that my going for a single drive in a fossil fuel car belongs to a large set of actions that causes climate change and its related harms. However, it is far from obvious how to establish membership of this intuitive sort. One quite straightforward idea is as follows: all acts that harm the victim belong to the relevant set in HARMLESS TORTURERS, and all acts that cause climate change and its related harms belong to the relevant set in the climate case. However, if counterfactual difference-making is what matters, as (C6) says, this solution is ruled out in collective impact cases. In these cases, no particular act harms anyone, and no particular act causes the collective outcome. This was the whole point of appealing to (C7) to begin with. If we could establish that each flipping of a switch in cases like HARMLESS TORTURERS was harmful, and similarly that each drive in a fossil fuel powered car is a cause of climate change and its related harms, (C7) would be otiose. We need some other way to single out the acts that belong to the relevant sets. There are many potential ways to do this, none of which succeeds.

First, we could appeal to intentions, as Kutz (2000) does. We could say that you have a reason not to aim for collective harm in acting, and that you likewise have a reason to aim for collective benefits. This is true most of the time, but appealing to intentions does not solve the current problem. In many collective impact cases, people do not aim for the collective outcome. These outcomes are merely the by-products of people doing ordinary things, aiming for something completely different. Climate change, for instance, is the by-product of people aiming to get around efficiently, or farmers growing rice to put food on the table, and so on. For more on this point, I refer the reader back to the discussion of Kutz's complicity principle earlier in this chapter.

Second, we could appeal to types of act. This seems to be Nefsky's (2017) preferred solution. The idea is that you have a reason not to perform an act of a certain type if it is true that, when enough such acts are performed, some collective harm will occur. And similarly, that you have a reason to perform an act of a certain type if it is true that, when enough such acts are performed, some collective benefit will occur. However, in adopting this route, we quickly run into the problem of how to describe the relevant types of act. Do I, for instance, have a climate-change-related

reason not to use a steam engine car running on fossil fuel? If we take the relevant act description to be “using a steam car” it will turn out that I do not. There is no way that climate change will be brought about as a result of enough people using steam cars. However, if we take the relevant act description to be “using a fossil fuel powered car”, it will turn out that I do have a climate-change-related reason not to use my car. It is quite likely that future climate-change-related harms will be brought about as a result of too many people using fossil fuel powered cars. It seems hard to find a principled way to identify the relevant act descriptions other than by saying that a relevant act description is any act description that accurately picks out acts that are causally connected to the outcome in a relevant way. However, if this is our solution, we are presupposing an account of what it is to be relevantly causally connected to the outcome. That is, if we say this, we run into the superfluity problem. I will say more about this in Chapter 4.

Third, we could try to draw a boundary around the relevant set of acts by saying that this is the smallest set of acts of which it is true that had none of these acts occurred, harm would not have occurred. This is Parfit’s (1984) solution. He argues that the relevant set of actions cannot be just any set of actions that together harms or benefits people. There are many sets of acts, and many of these sets harm people, or benefit people, and any of our acts might belong to such a set. For illustration, he considers the following case (here reformulated and abridged):

ASSASSINS WITH FRED ASTAIRE: Two assassins simultaneously shoot me, and each shot is sufficient for my death. Meanwhile, happily unaware, Fred Astaire is dancing in the distance.

Parfit argues that the relevant set must be the one consisting of the two shots, not the one consisting of the two shots and Fred Astaire’s dancing in the distance. Still, according to (C6) this larger set also harms me. Had none of the elements in this larger set occurred, I would have survived. In turn, if we say, *per* (C7), that an act might be wrong if it belongs to a set of acts that harms someone, we will have to conclude that Fred Astaire might have acted wrongly when unknowingly dancing in the distance. This conclusion is absurd, says Parfit. To avoid including acts that are completely unconnected with the outcome in the relevant sets, he suggests that we adopt the following principle:

(C8) When some group together harm or benefit other people, this group is the smallest group of whom it is true that, if they had all acted differently, the other people would not have been harmed, or benefited.

(Parfit 1984: 71)

This principle does the trick in ASSASSINS WITH FRED ASTAIRE. The smallest group of whom it is true that if they had all acted differently I would have survived does not include Fred Astaire. There is a smaller such group, namely the one consisting only of the two assassins.

However, as Gruzalski (1986) points out, (C8) is not always accurate. In non-threshold cases, there is no smallest group of whom it is true that, if they had all acted differently, the outcome would not have occurred.<sup>39</sup> In HARMLESS TORTURERS, for instance, you could take any group of which it is true that, if they had all refrained from flipping their switch, the victim would not have been harmed, and remove one of the harmless torturers from this group, and still wind up with a group of which it is true that, if they had all refrained from flipping their switch, the victim would not have been harmed. This is so because no single harmless torturer makes a difference to what the victim feels.

As a consequence, using (C8), we could never establish that a particular harmless torturer belongs to the relevant group. Any given torturer could argue “It is they who harm the victim. Not me”. The deeper problem, however, is that there is no smallest group of whom it is true that, if they had all acted differently, harm would not have occurred. Since there is no such group, (C8) fails to pinpoint any group at all that harms the victim.

These points generalise to all non-threshold cases. Thus, any driver equipped with Parfit’s principles could reason “My drive makes no difference to climate-change and its related harms, so I have no climate-change-related reason not to drive. The problem is that there are so many other drivers that use fossil fuel cars, not that I use fossil fuel”. He could add that there is in fact no group that causes climate change and its related harms. These points affect Schwenkenbecher’s argument.

Parfit’s principles also give the wrong verdicts in some threshold cases. Consider the following case:

THE LAKE: You, Vanessa and Walter all live close to a lake with a sensitive ecosystem. Each of you has a boat. If two or more of you paint the hulls of your boats with a cheap and toxic paint rather than a more expensive but non-toxic alternative, the ecosystem in the lake will collapse. If at most one of you uses the toxic paint, the ecosystem will continue to thrive. As it turns out, all three of you use the cheaper paint, and the lake becomes a wet wasteland.

(Adapted from Björnsson 2014)

---

<sup>39</sup> Gruzalski comments on Parfit’s (1984) principles (C10), (C11) and (C12), but the same issue can be raised about (C8).

In this case, there is no *smallest group* of which it is true that, if all its members had acted differently, the ecosystem in the lake would not have collapsed. It is indeterminate which group this is. It could be any of the following groups: you and Vanessa, you and Walter, or Vanessa and Walter. The problem is slightly different here from that in non-threshold cases. In the latter, there is no smallest group at all. In threshold cases like THE LAKE, there is a tie between different candidates for being the smallest group, meaning that it is indeterminate which group is the smallest one. This is just as bad as not pinpointing any group at all. Intuitively, it is clear that all three of you caused the ecosystem to collapse.

Moreover, just as any individual driver believing in Parfit's principles could blame the other drivers for bringing about climate change, any of the boat owners in THE LAKE who likewise embraces these principles could defend herself by saying: "It is those other guys that caused the ecosystem to collapse, not me. What I did didn't make any difference to the survival of the ecosystem!". Unless we replace or revise (C8), we have no principled way to counter this argument. So, (C8) gives counterintuitive verdicts in a range of cases.

Fourth, Jan Willem Wieland and Rutger van Oeveren (2020) have recently suggested that "one is part of the relevant group *when one adds to the given underlying dimension*" (180). One does that – adds to the underlying dimension of some outcome – where it is the case that this outcome will occur if enough such additions are made.<sup>40</sup>

More precisely, they say that:

(P2) S is a member of the set that causes O iff S does the act X which is such that: (i) X adds to an underlying dimension D, and (ii) because enough others add to D, D causes O.

(Wieland & van Oeveren 2020: 177)

Consider HARMLESS TORTURERS. The thought is that you are part of the group that causes the victim harm here if and only if: you flip your switch, and in doing so, (i) you increase the voltage going through the victim, and (ii) the victim is in pain because enough others likewise increase the voltage going through him. Similarly, you are part of the group that causes climate change if and only if: you go for a drive in a fossil fuel car, and in doing so, (i) you increase the amount of greenhouse gas in the atmosphere, and (ii) climate change and its related harms occur because enough others similarly add greenhouse gases to the atmosphere.<sup>41</sup>

---

<sup>40</sup> When identifying the underlying dimension, they refer to Kagan (2011) and Nefsky (2012).

<sup>41</sup> According to Wieland and van Oeveren, you have a reason not to perform some act if doing so makes you a member of a set that causes harm. This makes me wonder why you would have a

The intuitive idea behind this proposal is that it is possible for you to contribute to some outcome even though this outcome would occur whether or not you contribute to it. I sympathise with this idea. This proposal also gives the right verdict in non-threshold cases, where Parfit's principle (C8), for example, fails. So far so good. The problem is that, in some cases, the proposal does not seem to capture what it is to contribute to an outcome (in a sense that is relevant to what reasons you have). To see this, consider the following scenario, proposed by Nefsky during the workshop "Small Acts, Big Harms" in Helsinki 2021:

VENDING MACHINE: A, B, and C are walking in a national park, when they come across two hikers who have been lost for days in the backcountry. They are starving. Luckily, there is a vending machine nearby, selling granola bars for \$4 each. The machine accepts all coins and bills, but it does not give change. The two starving hikers do not have any money. But A has a \$5 bill, B has a \$10 bill, and C has a quarter. There is no one else around.

(Nefsky 2021)

Here, the intuitive verdict is that C has no reason to put his quarter into the vending machine. There is no way his doing so would contribute to the hikers' getting something to eat.<sup>42</sup> However, this is not the verdict that Wieland and van Oeveren's account gives. To add a quarter to the vending machine would be to add to the underlying dimension in this case. It is an instance of adding money to the machine, and when enough money is added, the hikers will get their granola bars. So (i) is satisfied. (ii) is also satisfied: the beneficial outcome will occur as a result of enough people adding money to the vending machine. So, the account Wieland and van Oeveren offer – that you have a reason to perform acts making you a member of a set that benefits others – entails the intuitively incorrect result that C has a reason to add a quarter to the vending machine.

As Nefsky sees it, putting a quarter into the vending machine is a superfluous part of the cause of the beneficial outcome. I would say that putting a quarter into the vending machine is not even a part of the cause of the beneficial outcome. Either

---

reason not to go for a second drive in a fossil fuel powered car. If you have already gone for such a drive, you already are a member of the relevant set. I think, however, that this particular problem can be avoided by instead saying that you have a reason not to perform an act that belongs to the set of acts that causes harm and then defining what it is for an act to belong to a set that causes harm along the lines of (P2).

<sup>42</sup> You might be tempted to think that C does have a reason to add a quarter. For instance, you may think there might be other people around who could add their quarters, and that this eventually would produce something to eat for the hikers. I agree that if there is a possibility that adding a quarter to the vending machine could make a difference, it seems that C has a reason to add a quarter. However, in VENDING MACHINE, there is no one else around, so putting a quarter into the slot is an entirely useless thing to do. It does not contribute to anything.

way, the problem is that Wieland and van Oeveren's principle fails to distinguish the morally relevant acts from the morally irrelevant ones. It therefore delivers mistaken verdicts about the reasons we have in collective impact cases.

There are other possible ways of deciding which acts belong to the set of acts that causes some collective outcome in collective impact cases.<sup>43</sup> I cannot consider them all here. I suspect that the only successful way of picking out the relevant acts is to identify a relevant causal link that connects each act to the collective outcome. Unless our account describes a relevant causal connection between each particular act and the outcome, it will either presuppose that there is such a connection, or give mistaken verdicts about some cases. That is, it will run into the superfluity problem or the disconnect problem.

### **Is the Set Harmful?**

It might seem obvious that the set of all flippings of switches harms the victim in HARMLESS TORTURERS. However, as Petersson (2004, 2018) points out, this is not the result we get if we apply (C6). According to this principle, "an act harms someone if its consequence is that someone is harmed more" (Parfit, 1984, 69). I take it that this principle applies to act sets as well as acts, and that "more" implicitly means "more than if the act had not been performed or the set of acts had not occurred". (C6), then, tells us to compare what happens in the actual world with what would have happened if the act or act set had had not occurred. If someone is harmed or benefitted more in the counterfactual world, the act or set harms/benefits this someone. What would have happened if not all the harmless torturers had flipped their switches? Most likely, all but one would have still flipped their switches, and the victim would still be in excruciating pain. However, in the counterfactual world where all but one harmless torturer flips their switches, the victim feels as much pain as he does in the actual world where all the harmless torturers flip their switches. By hypothesis, one flipped switch does not make any difference to how the victim feels. So, by (C6), the set of all flippings of switches

---

<sup>43</sup> Young (2011), for instance, does say that everyone who participates in an unjust structural process has a forward-looking responsibility to offer redress for this process together with others. For illustration, everyone participating in the global garment industry has a forward-looking responsibility to provide redress for the injustices within this industry – e.g. the unjust working conditions under which many of our clothes are produced. As you see, Young is aiming to explain, not the intuition that you have a reason to refrain from performing some act in collective harm cases, but rather the thought that you have a reason to alleviate the collective harm in some other way. Still, you might think that we could use her account of what it is to participate in an unjust structural process to find out who belongs to a particular harmful set. One problem with this idea, however, is that Young says little or nothing about how to distinguish participants from non-participants. We have to extract a principled way of making that distinction by considering the examples she offers. This requires some elaborate exegesis. I address Young's account in Gunnemyr (2020).



does not harm the victim. He is not harmed more in the closest possible world where this set does not occur.

The same point can be made in terms of causation. It seems obvious that the set of all drives with fossil fuel cars causes climate change and its related harms. However, if we assume a simple counterfactual account of causation on which C causes E if, and only if, E would not have occurred if C had not occurred, we get the result that the set of all drives with fossil fuel cars does not cause global warming. Why? The closest possible world where this set does not occur is a world where all but one of the drives occur. If we assume that a single drive in a fossil fuel car does not make any difference to climate change, we find that there is no difference between the actual world and the closest possible world in terms of climate-change-related harms. So, contrary to expectation, it turns out that the set of all drives does not cause any climate-change-related harms.

Here, you might be tempted to argue that we have not made the relevant comparison. You might suggest that we should take our cue from (C8) also when applying (C6), and compare what happens in the actual world with what would have happened if *no* harmless torturers had flipped their switches, and again that we should compare what happens in the actual world with the world where *no one* from now on drives fossil fuel cars. That would entail the results we want – that the set of all harmless torturers *does* harm the victim, and that the set of all drives with fossil fuel cars *does* cause climate-change-related harms. The problem with this line of argument is that it is unclear why we should consider worlds that are further away from the actual one. Why are the relevant counterfactual worlds those where no harmless torturers flip their switches and no one uses fossil fuel cars? Parfit does not say anything about this issue. So, at best, we could say that it is indeterminate on Parfit's account which world we should consider, and so indeterminate whether these sets harm other people. However, this result is also problematic. Intuitively, it seems that the set of all flippings of switches in HARMLESS TORTURERS does harm the victim, and that the set of all drives in fossil fuel powered cars *is* a cause of climate change and its related harms.

In contrast, Lewis (1973a, 1973b) argues that the relevant counterfactual world for causation is the closest possible world where the relevant event does not occur. This is now the standard position in the causation literature. If we approve it, and if we accept a simple counterfactual account of causation, we must conclude that the set of all drives with fossil fuel cars does not cause climate change and its related harms.

At this point I think we should say, first, "So much the worse for the simple counterfactual account of causation!" If we abandon this account, we can argue that a single flipping of a switch is a cause (one of many) of the victim's pain, and that a single drive is a cause (one of many) of climate change and its related harms. We could also argue that the set of all flippings of switches is a cause of the victim's pain, and that the set of all drives in fossil fuel powered cars is a cause of climate

change and its related harms: If these sets had not occurred, some similar sets would have occurred instead, but these other sets would be relevantly different from the actual ones. They would be one cause short of the actual sets.

I also think we should say, second, “So much worse for the idea that the relevant comparison in (C6) is with what would have happened if the act or act set had not occurred!” It seems that you might harm someone even when this harm would have occurred if you had acted otherwise. However, I will not say much more about what it is to harm or benefit someone in this thesis. That will have to be the topic for another day.

## Indirect Consequences

Maybe we can explain the reasons intuition by appealing to indirect consequences. For instance, if I pour my pint into the cart in DROPS OF WATER, I might influence others to likewise pour their pints into the cart. Conversely, if I do not pour my pint into the cart, I might equally influence others to do the same. Similarly, by refraining from joy-guzzling, I might encourage others to reduce their emissions, and if I do go for a leisure drive, others, seeing this, may do the same. As Jamieson argues, “One’s behavior in producing and consuming is important [...] for the example-setting and role-modeling dimensions of the behavior” (2007: 179). Holly Lawford-Smith (2015) and Melissa Lane (2018) have recently proposed similar ideas.

The appeal to indirect consequences is really a causal response to the inefficacy problem, so it would, perhaps, be better taken up in the next chapter. However, this response runs into the same problems as the non-causal ones, and therefore I discuss it here.

Sinnott-Armstrong (2005) argues that the appeal to indirect consequences cannot solve the inefficacy problem. His argument goes as follows: If it is true that each act is ineffectual, and if I influence others to refrain from joy-guzzling by refraining myself, I only influence others to perform a number of ineffectual acts. We might then ask: why would I have a reason to influence others to perform ineffectual acts? Unless we can explain in what way a particular joyride is not completely ineffectual, it does not matter whether I influence others to act like me.

Jamieson and others could respond that while it is true that each individual act that I encourage is ineffectual, a sufficient number of them makes a difference. So, my reason not to joy-guzzle, if I have one of this kind, is that my refraining from doing so influences others, in sufficient numbers, to follow my example. This brings us to what I think are the two basic problems of appealing to indirect consequences when trying to explain the reasons intuition.

First, we run into the inefficacy problem also when considering how our acts might influence others. Typically, people's attitudes to joy-guzzling and the general use of fossil fuel cars will be the same whether or not I refrain from joy-guzzling at one particular occasion. Plausibly, people will be influenced only if enough others refrain from joy-guzzling, wear t-shirts saying "Don't joy-guzzle!", and so on. In addition, even if someone were to explain that my avoidance of joy-guzzling triggered them to reduce their own, it would be quite likely that they would have reduced their joy-guzzling anyway. My restraint was probably only the last straw. If I had not refrained, something else would soon have made them reduce their joy-guzzling. So, unless we show what is wrong with the inefficacy argument, it seems that anyone could argue: "My joy-guzzling this Sunday won't make anyone less inclined to use fossil fuel cars, so I might as well go ahead with it!" It seems, then, that the appeal to indirect consequences only works if we can show that there is some relevant connection between refraining from joy-guzzling and others' being influenced besides counterfactual difference-making. In other words, this appeal runs into the superfluity problem.

Second, as Nefsky (2019) argues, even if we agree that I have a climate-change-related reason to influence others to reduce their emissions – perhaps we think that there is a tiny chance that I could make some difference to whether they reduce their emissions – I could do so in ways other than by refraining from joy-guzzling. I could, for instance, wear a t-shirt saying "Don't joy-guzzle!" So, even if I have a reason to influence others, this is not specifically a reason to refrain from joy-guzzling. Additionally, we can think of cases where I influence nobody by going for a leisure drive (no one notices). However, it seems that I would have a climate-change-related reason not to joy-guzzle also on these occasions. Thus, the appeal to indirect consequences as a response to the inefficacy argument runs into the disconnect problem. It does not explain why I have a reason specifically not to joy-guzzle, and it fails to unearth some climate-change-related reasons I do seem to have.

## Conclusion

This concludes my discussion of non-causal responses to the inefficacy argument. I have explained that these responses run into either the disconnect problem or the superfluity problem. When they run into the disconnect problem, they identify some reason other than the outcome-related reasons we are looking for. For instance, they disclose reasons to be fair, or reasons to take collective action. These are important reasons, but they do not explain where the inefficacy argument goes wrong. When, instead, they run into the superfluity problem, the responses deliver the outcome-related reasons we are looking for but only by assuming, illicitly, that there is a relevant causal connection between the relevant  $\phi$ -ing and the collective outcome.

You might think my aim has been to prove that the theories I have discussed (virtue ethics, Kantianism, and so on) are mistaken. This was not my intention, and I do not take my arguments to show this. My arguments merely show that it is most likely that we cannot appeal to fairness, virtue ethics, Kantianism, complicity, reasons to take collective action, membership of a group or indirect consequences in order to fully explain what is wrong with the inefficacy argument. In collective impact cases, these appeals only succeed when we assume, or show, that there is, a causal connection between  $\phi$ -ing and the collective outcome. Later (in Chapter 5), I will show that there is such a connection, and hence provide these theories with a basis on which they can, in principle, respond to the inefficacy argument.

I have not established that no non-causal response could explain the reasons intuition without running into the disconnect or superfluity problem. I have not considered every possible non-causal response. There is, for instance, a wide variety of virtue ethics theories which I have not discussed, and more can be said about Kant's formulas. Maybe one of these theories is capable of explaining the reasons intuition. I suspect, however, that they will all be thwarted by the problems very like those we have been reviewing. If we want to solve the inefficacy problem, it seems more promising to ask whether we can isolate a relevant causal connection between  $\phi$ -ing and the collective outcome.

There have been hints that this causal connection is one of contribution. If we could explain how, exactly, pouring a pint into the cart contributes to the alleviation of suffering in DROPS OF WATER, Cullity's (2000) point that I would "be relying on others to do what we ought collectively to be doing, without contributing [myself]" (15) if I do not pour my pint into the cart would gain traction. Similarly, FUL could then explain why I have reason to donate my pint. I have an imperfect duty to sometimes help others, and if pouring my pint into the cart contributes to the alleviation of suffering, this seems to count as helping others. Again, Lawson's (2013) suggestion that "I am accountable for what others do when I knowingly contribute to a harmful outcome that results from our collective contributions" (234) seems to be on the right lines if we can explain how acting in the relevant way contributes to the harmful collective outcome.

So far, I have not said much about the notion of contribution. I have relied on an intuitive grasp of it. Briefly, I take contribution to be a causal notion. If an act contributes to an outcome, there is some causal connection between the act and the outcome. I will leave it unsaid for now exactly what I think this causal connection consists in, but candidate accounts include the following: When an act contributes to an outcome, it makes a counterfactual difference to the outcome, or raises the probability of the outcome, or increases the security of the outcome, or makes the outcome occur sooner rather than later. There are doubtless other ways of explaining contribution.

In what follows, then, the notion of contribution is going to play a central role. I will consider whether David Lewis' (1973a, 1986a, 1986b) early account of causation, or Richard Wright's (1985, 2013) NESS condition of causation, describes the relevant causal connection between  $\varphi$ -ing and the collective outcome that we are looking for, and argue that they do not. My claim will be that they fail to describe the relevant causal connection exactly because they fail to capture the idea that a cause must contribute to its outcome.

### 3. Causal Responses

Do you cause climate change and its related harms if you go for a leisure drive in a fossil fuel powered car? In general, do you cause the relevant outcome if you  $\varphi$  in collective impact cases? Quite a few philosophers frown upon this idea. They typically argue that since the outcome will occur whether you  $\varphi$  or not in the relevant cases,  $\varphi$ -ing does not cause the outcome. For instance, Walter Sinnott-Armstrong (2005) says that a single drive “does not cause global warming, climate change, or any of their resulting harms” (299). Why? The answer is, because such a drive is neither necessary nor sufficient for harm to occur. Dale Jamieson (2007) similarly argues that “Joyriding in my ‘57 Chevy will not in itself change the climate, nor will my refraining from driving stabilize the climate” (167), concluding that climate change and its related harms are not a consequence of any single drive.

Christopher Kutz (2000) makes a similar point. He argues that THE INDIVIDUAL DIFFERENCE PRINCIPLE – which says that “I am accountable for a harm only if what I have done made a difference to that harm’s occurrence” (116) – entails that I am not accountable for the outcome in cases of overdetermination. For illustration, he considers the Allied strategic bombing of Dresden in 1945 (we referred to this in the previous chapter). As Kutz describes these bombings, they were inhumane and strategically useless. The bombing and the resulting firestorm killed approximately 20,000 civilians and devastated the city centre. There were more than a thousand planes involved and around eight thousand crewmembers directly participated as pilots, gunners, navigators and bombers. Since the outcome was overdetermined, Kutz argues, no individual crewmember can be said to have caused it, so none cannot be morally responsible for it – at least, not according to THE INDIVIDUAL DIFFERENCE PRINCIPLE. As Kutz puts it, “No rank-and-file individual made a difference to the evil that occurred”, and therefore, each “bomber can truly reply to the victims or their survivors, ‘Why blame me? I have not caused your suffering’” (2000: 122).<sup>1</sup> Elizabeth Cripps (2013) similarly suggests that climate change “can effectively be treated as overdetermined” (123), at least if there are more emitters than needed to trigger any given climate-change-related harm.<sup>2</sup>

---

<sup>1</sup> As pointed out earlier, he makes the same point in relation to environmental harms.

<sup>2</sup> Jackson (1997) gives a similar argument, albeit in relation to whether one particular act in a collective harm cases is *harmful*. Considering ASSASSINS, where two assassins simultaneously shoot me, and where each shot was sufficient for my death, he contends that while the outcome is

These arguments presuppose something like a simple counterfactual analysis of causation (or a but-for condition, as it is sometimes called in legal philosophy) which states that a cause is necessary for its effect. That is, they presuppose something like:

SIMPLE: C causes E if and only if: had C not occurred, E would not have occurred.<sup>3</sup>

(Differently put: C causes E if and only if E *counterfactually depends* on C.)

Now the question is: Can we really assume that SIMPLE gives correct verdicts about causation in collective harm cases? It seems that we cannot. In collective harm cases, no individual act makes a difference to the outcome, and SIMPLE is well-known for giving counterintuitive results in precisely these kinds of case. Consider, for instance, the following pre-emption case:

WINDOW BREAKING: Suzy throws a rock at a window, breaking it. If Suzy had not thrown her rock, Billy would have thrown a rock a moment later, and the window would still have broken. As things proceed, however, Billy never throws his rock.<sup>4</sup>

Here, the intuitive verdict is that Suzy's throwing her rock caused the window to break. Still, the window shattering does not counterfactually depend on what Suzy did. Had she not thrown her rock, Billy would have thrown his, and the window would have broken anyway. So SIMPLE does not entail that Suzy caused the window breaking.

That SIMPLE entails that Suzy did not cause the window to break is not usually taken as a sign that, in fact, she did not cause the window to break. Rather, it is taken to show that there is something wrong with SIMPLE. Likewise, it seems that if SIMPLE entails that a leisure drive in a fossil fuel powered car is not a cause of climate change, we should not – at least, not straightforwardly – take this as a sign that such a drive does not cause climate change. Instead, we should pause and ask ourselves what the relevant notion of causation is. As Björn Petersson (2013) says:

---

unfortunate, neither assassin harmed me. Jackson does not couch his discussion in terms of causation, and so does not commit himself to any particular view of causation, but the general idea is the same: unless your action makes a difference to the outcome, there is no morally relevant connection between what you do and this outcome.

<sup>3</sup> Typically, proponents of a simple counterfactual account of causation (or some elaborated version of it) also presuppose that E occurs after C. Without this presupposition, we would run into problems with backward causation.

<sup>4</sup> Cases like this are standard in the literature on causation. See e.g. Lewis (2000) and Hall (2004).

Kutz and others oversimplify the relation between counterfactual dependence and causation, and they overlook the possibility that causal relations other than marginal contribution could be morally relevant.

(Petersson 2013: 849)

Maybe there is some morally relevant causal relation that could explain, for instance, why I have a climate-change-related reason to refrain from going for a joyride in a gas-guzzling car.

While there is a rich variety of theories of causation (e.g. Salmon 1994; Dowe 2000; Hitchcock 2001; Woodward 2003; Hall 2004; Schaffer 2005), it has been taken more or less for granted in the literature that counterfactual dependence is the only morally relevant causal relation. There are a few exceptions, however. Recently, Anton Eriksson (2019) has argued that David Lewis' (1973a, 1986b, 1986a) early account of causation captures the required relation, and Matthew Braham and Martin Van Hees (2012) have suggested that Richard Wright's NESS condition does so. (Wright would certainly agree.) In the rest of this chapter, I will argue that both Lewis' early account of causation and Wright's NESS condition fail to describe the morally relevant causal connection we are looking for. In essence, they fail because you only have an outcome-related reason to act in a certain way if acting in this way *contributes* to the outcome, yet these theories entail that you can cause an outcome without contributing to it. (This point will become clearer soon, I hope.) Still, even though neither Lewis nor Wright provides a causal analysis that accurately helps us to distinguish situations where we have outcome-related reasons from situations where we do not, there might be other causal analyses that do. Exploiting this possibility, Chapter 5 argues that the relevant causal relation is security-dependence.

Before considering Lewis' and Wright's accounts in detail, I need to make some preliminary points. First, Sinnott-Armstrong (2005) does not simply argue that a single drive does not cause climate change and its related harms because a cause is necessary for the occurrence of its effect while a single drive is not necessary for the occurrence of climate change (etc.). He makes the more intricate argument that my drive does not cause climate change (etc.) because my drive was *neither necessary nor sufficient* for climate change (etc.) to occur. Unfortunately, he does not further explain what account of causation he has in mind. A fair guess is that his point is something like this: Regardless of whether the correct account of causation is one according to which a cause is necessary for its effect (such as SIMPLE) or one according to which a cause is sufficient for its effect (as is claimed in NESS), we have to conclude that my drive does not cause global warming.<sup>5</sup> If what I say in this

---

<sup>5</sup> Sinnott-Armstrong (2005) also discusses the possible objection that, sometimes, it seems that some event causes another *even if* it is neither necessary nor sufficient for its effect. This might happen, he argues, if this event stands out as particularly salient, which it does if it was intentional or rare. For discussion, see e.g. Gunnemyr (2019). Here, I set this discussion aside.



chapter and the following chapters is correct, Sinnott-Armstrong may be right that a single drive neither is necessary nor sufficient for climate change and its related harms to occur, but wrong that there is no morally relevant causal connection between a single drive and climate change.

Second, Kutz (2000) does not unreflectively assume that what I call SIMPLE is the only morally relevant notion of causation. He considers Martin Bunzl's (1979) suggestion that we should individuate events by their causal antecedents in order to better handle cases of overdetermination. When we do, it turns out that the Dresden bombings would not have occurred if one of the bombers had decided not to drop his bombs. Instead, a slightly different event would have occurred – an event with slightly different causal antecedents. Kutz agrees that Bunzl's solution might solve the metaphysical problem of who causes what in cases of overdetermination. However, he continues, mere metaphysical differences do not have to be morally relevant. As he puts it, “we can stipulate solutions to the relevant metaphysical causal riddles without illuminating the ethical questions at all” (125). To see why Bunzl's proposal cannot help us explain why each crewmate had a reason not to participate in the bombings, Kutz asks us to consider a potential bomber who refrained from participating in the Dresden bombings. If he had not refrained, the bombings would have had different causal antecedents. So, on Bunzl's proposal, the crewmate's refusal to participate was a cause of the bombings that actually occurred. If we combine this idea with the idea that you are accountable for the harm you cause, it turns out that the potential crewmember who refused to participate in the bombings was accountable for them. This verdict, Kutz argues, must be mistaken. I agree. However, this does not entail, as Kutz seems to think, that SIMPLE describes the only morally relevant causal connection. There could be other kinds of causal relations that matter morally.

Finally, in the background of this discussion, there is an assumption that you have a reason not to cause harm to others.

THE HARM PRINCIPLE: You have a reason not to perform an act that causes harm to others.<sup>6</sup>

This principle explains why, for instance, you have a climate-change-related reason not to go for a drive in a gas-guzzling car if doing so causes climate-change-related harm to others.

Harm can here be understood in at least two ways. It could be taken to be a bad state of affairs. On this view, if you cause harm to others, you cause others to be in a bad state of affairs. If you cause climate-change-related harms to others, you might, for

---

<sup>6</sup> For early formulations of this principle, see Mill (2008/1859), Ross (2002/1930) and Feinberg (1984).

instance, cause others to have their houses flooded, or suffer a shortage of food. This view of harm might be called a *non-comparative* view of harm. This view is well-known for giving counterintuitive results in some cases, most importantly perhaps in cases where you make someone better off than she used to be by putting her in a bad state of affairs. Here, it seems that you helped her rather than harmed her, yet this is not what a non-comparative view of harm entails.

To avoid this problem, those who write on harm have tended to favour a *comparative* view of it. According to the standard version of his view – the *counterfactual comparative* view of harm – an event harms someone if and only if this person would have been better off had this event not occurred. Climate change, for instance, harms people, because some people would have been better off if it had not occurred. For example, they might have been better off if a climate-change-related flood or drought had not occurred. Climate change might also benefit people. That is, there may be people who would have been worse off if it had not occurred. Still, we shall assume that climate change, overall, is a harmful event: it harms people more than it benefits them. Anyway, if THE HARM PRINCIPLE holds, the mere fact that climate change harms people gives us reasons not to perform acts that cause climate change, reasons that might or might not be outweighed by other considerations. In what follows, it does not matter whether you have a non-comparative or a comparative counterfactual view of harm. The arguments apply just the same.<sup>7</sup>

Commonly, THE HARM PRINCIPLE is stated in terms of moral obligations (Sinnott-Armstrong 2005; Hiller 2011; Kingston & Sinnott-Armstrong 2018; Eriksson 2019). As Sinnott-Armstrong (& Kingston) understand this principle, a moral obligation is something you must do without exception or qualification. However, if the principle is interpreted in this way, it is obviously mistaken. Surely, for instance, we are justified in brusquely pushing another person aside if we need to do so to save a drowning child. THE HARM PRINCIPLE, however, implies that brusquely pushing someone aside is something you must not do, even in order to save a drowning child. Avram Hiller (2011) suggests that we ought instead to specify THE HARM PRINCIPLE as one stating we have a *pro tanto* moral obligation not to perform an act that causes harm to others, where a *pro tanto* moral obligation is, very roughly, a strong, but overridable, reason. This seems correct to me. This is also roughly how Eriksson understands the principle. Here, however, I propose to consider an even less demanding version of THE HARM PRINCIPLE – one stating simply that you have a reason (strong or not) not to cause harm to others. I do this because it makes explicit the connection with the inefficacy problem, which is stated in terms of reasons.

---

<sup>7</sup> For illuminating discussions on how to understand what it is to harm someone, see Algander (2013), Feit (2015), Petersson (2018) and Johansson and Risberg (2019).

## Lewis' Simple Account

Eriksson (2019) suggests that Lewis' (1973a, 1986a, 1986b) early account of causation describes a causal relation that is relevant to the reasons we have. If we follow Lewis, Eriksson argues, we can explain the intuition that we have climate-change-related reasons not to joy-guzzle. Such a drive might be a cause of climate-change-related harms, and we have reasons not to cause harms. It seems to me that he is correct that a single leisure drive might be a cause of climate-change-related harms, but that he is mistaken in thinking that Lewis' early account describes a causal relation that is relevant to the question which outcome-related reasons we have.

Here, I will first introduce Lewis' early account of causation and Eriksson's argument. I will then explain why I think we cannot use Lewis' early account to show that we have reasons not to go for a leisure drive.

### Lewis' Early Account

Lewis (1973a, 1986a, 1986b) takes causation to be the ancestral of counterfactual dependence, where counterfactual dependence is understood along the lines of SIMPLE, as follows:

SIMPLE WITH TRANSITIVITY: C causes E if and only if there is a causal chain leading from C to E, where a causal chain is a sequence of counterfactual dependencies.

He appeals to the transitivity of counterfactual dependence to handle cases of pre-emption. As we saw, SIMPLE has the counterintuitive implication that Suzy did not cause the window breaking in WINDOW BREAKING. SIMPLE WITH TRANSITIVITY, however, gives the intuitively correct verdict that Suzy did cause the breakage. Say that C is Suzy's throwing her rock, and E the window breaking. Now, consider the intermediate event D which consists in Suzy's rock flying through the air towards the window. Here, D counterfactually depends on Suzy's throwing the rock. Had Suzy not thrown the rock, it would not be flying through the air. Moreover, the window breaking counterfactually depends on D. If the rock had not been flying towards it, the window would not have broken. So, there is a sequence of counterfactual dependencies leading from Suzy throwing her rock, via the rock flying through the air, to the window breaking. Given this, SIMPLE WITH TRANSITIVITY entails that Suzy throwing her rock caused the window to break, and that this is so even though the window breaking did not directly counterfactually depend on Suzy throwing her rock.

WINDOW BREAKING is a case of *early* pre-emption. In such cases, there is some intermediate event *D* that counterfactually depends on *C*, and that is necessary for the outcome. In WINDOW BREAKING, for instance, the intermediate event consisting in the rock flying towards the window counterfactually depends on Suzy throwing a rock, and is necessary for the window to break. It is necessary for the window to break because, at this later point, Suzy has already prevented Billy from throwing his rock by throwing hers.

While SIMPLE WITH TRANSITIVITY gives the right verdict in early pre-emption cases, it gives counterintuitive verdicts in *late* pre-emption cases. In the latter, there is no intermediate event *D* that counterfactually depends on *C* and also is necessary for the outcome *E*. In Paul and Hall's words: "at no point in the sequence of events leading from cause to effect does there fail to be a backup process sufficient to bring about that effect" (2013: 99). Therefore, the strategy we used to show that Suzy caused the window to break in WINDOW BREAKING does not work. Consider, for instance, the following late pre-emption cases:

[SHOOTING AND POISONING:] *D* shoots and kills *P* just as *P* was about to drink a cup of tea that was poisoned by *C*.

(Wright 1985: 1775)

and

[BOTTLE SHATTERING:] Billy and Suzy throw rocks at a bottle. Suzy throws first, or maybe she throws harder. Her rock arrives first. The bottle shatters. When Billy's rock gets to where the bottle used to be, there is nothing there but flying shards of glass. Without Suzy's throw, the impact of Billy's rock on the intact bottle would have been one of the final steps in the causal chain from Billy's throw to the shattering of the bottle. But, thanks to Suzy's preempting throw, that impact never happens.

(Lewis 2000: 184)

Intuitively, *D*'s shot caused *P*'s death and Suzy's throw caused the bottle to shatter. However, this is not the verdict SIMPLE WITH TRANSITIVITY yields. (Nor is it what SIMPLE says.) There is no event occurring after *D* took his shot and before *P* died which is necessary for *P*'s death. Whichever point in time we consider within this time interval, *P* will die whether or not *D* shoots him. That *P* is about to drink the poisoned tea guarantees this outcome. Likewise, there is no event occurring after Suzy threw her rock but before the bottle shattering takes place that is necessary for the bottle to shatter. Whichever point in time we consider within this time interval, the bottle will shatter whether or not Suzy's rock had come flying towards the bottle. Billy's rock guarantees the outcome.

So, SIMPLE WITH TRANSITIVITY correctly entails that the pre-emptive cause (e.g. Suzy's throw in WINDOW BREAKING) is a cause of the outcome in early pre-emption cases, but incorrectly entails that the pre-emptive cause (e.g. D's shot, Suzy's throw in BOTTLE SHATTERING) is not a cause of the outcome in late pre-emption cases.

## Eriksson's Argument

To return to the question of climate change, could SIMPLE WITH TRANSITIVITY coupled with THE HARM PRINCIPLE explain why you have a reason not to go for a leisure drive in your fossil fuel powered car? Eriksson (2019) argues that this is the case.<sup>8</sup> He acknowledges that SIMPLE WITH TRANSITIVITY runs into trouble in late pre-emption cases. However, he argues, this does not matter in the issue at hand. We can still use SIMPLE WITH TRANSITIVITY to show, for instance, that buying a flight ticket or going for a leisure drive in a car causes harms.

To make his argument, Eriksson first distinguishes between *emissions generating actions* like driving a car or flying a plane over the Atlantic, and *actions in causal chains leading up to emissions generating events*, such as oil companies selling oil or someone's buying a flight ticket. Actions in causal chains leading up to emissions generating events, he argues, are almost always early pre-emptive causes; they stand in the same relation to the emissions generating events as Suzy's throwing her rock does to the window breaking in WINDOW BREAKING. Therefore, SIMPLE WITH TRANSITIVITY entails that such actions cause the later emissions generating events. Buying a flight ticket causes the flight over the Atlantic, and selling petrol to a car owner causes the ensuing emissions when this car owner goes for a Sunday drive.<sup>9</sup>

---

<sup>8</sup> He also argues that "the agents of supply chains cause emissions in virtue of their actions forming parts of sets of *joint causes* on which the emission of GHGs counterfactually depend" (Eriksson 2019: 3, my emphasis). On a superficial reading, this idea seems susceptible to the charge that there is no straightforward way of distinguishing the agents that belong to relevant set from those who do not. It may also seem to be susceptible to the charge that the set did not cause global warming. In the closest possible world where this set does not occur, a very similar set occurs (e.g. the same set minus one leisure drive) and climate change and its related harms would have occurred just the same. However, Eriksson avoids these problems, since he takes "a set of events  $\{c, d\}$  to be a joint cause of an event  $e$  iff there is a causal chain leading from each member of that set to  $e$ " (2019: 18), where a causal chain is understood as a sequence of counterfactual dependencies. So, in the end, on Eriksson's account, it is not whether you belong to the group that causes harm that matters, but whether you (together with others) cause an outcome, where causation is understood along the lines of SIMPLE WITH TRANSITIVITY.

<sup>9</sup> For the sake of argument, I will grant that actions in causal chains leading up to emissions generating events are almost always early pre-emptive causes. However, the case that Eriksson uses to show that they are seems to be a case of late pre-emption. This case, in shortened form, is as follows:

AIRLINE: A particular airline suspends any flight if fewer than 100 tickets are sold for this flight. Each flight can take substantially more than 100 passengers. For today's flight, 110

The question remains, however, whether the emissions generating actions cause harm. Here, Eriksson concedes that some emissions generating actions are cases of late pre-emption – they stand in the same relation to harms as Suzy’s throwing a rock does to the bottle breaking in BOTTLE SHATTERING – and that these therefore do not cause the harms. However, he argues, not *all* emissions generating actions are like this. At least some such actions stand in the same relation to climate-change-related harms as Suzy’s throw does to the window breaking in WINDOW BREAKING. Some emissions generating actions might even make a counterfactual difference to climate-change-related harms. Therefore, SIMPLE WITH TRANSITIVITY entails that some emissions generating actions cause climate-change-related harm, and THE HARM PRINCIPLE entails, in turn, that I have a reason not to go for a drive, buy a flight ticket, and so forth, since my doing so might cause harm to others.

That some emissions generating actions cause harm is particularly obvious when we recall that harms can be fragile events. In Lewis’ words, an event is fragile “if, or to the extent that, it could not have occurred at a different time, or in a different manner” (1986a: 196). For instance, if a particular drought would not have occurred if I had refrained from leisure driving, but a very similar drought would have occurred instead, my drive caused the drought that actually occurred. If we assume, reasonably, that the drought was harmful, we can conclude (via THE HARM PRINCIPLE) that I had a reason not to go for the leisure drive.

## Problems with Transitivity

I agree with Eriksson that we can explain why I have a reason to refrain from going for a Sunday drive in a gas-guzzling car by looking at causation. However, I do not agree that SIMPLE WITH TRANSITIVITY describes a causal relation that is relevant to what outcome-related reasons we have. This principle appeals to the transitivity of counterfactual dependence in order to handle cases of early pre-emption, but since

---

tickets are sold and the flight leaves as planned. You bought one of these tickets. In fact, you were the 73rd person to buy a ticket for this flight.

Here, there is no point in time when my purchase was necessary for the flight to leave as planned. The forthcoming additional 10 purchases (above the first 100) at each point in time guarantees this outcome. Just as Billy threw his rock in BOTTLE SHATTERING whether or not Suzy threw her rock, the additional ticket purchasers here buy their tickets whether or not I make my purchase. Still, the picture changes if we factor in that events can be fragile. My purchase did, for instance, make a difference to the more fragile event “the flight takes off as planned with me on it”. Had I not bought a flight ticket, that event would not have occurred. However, as I will argue shortly, we should be wary of bringing in fragile events. Eriksson’s discussion of this case is presented in his (2019: 30-32).

it appeals to transitivity, it cannot reliably distinguish contributors from counteractors.<sup>10</sup> To see why, consider the following case:

CAR KEYS: One Sunday morning, I hide my friend's car keys in the hope of making her come along for a bike ride instead of going joy-guzzling as she usually does. However, she manages to hot-wire her car, and goes joy-guzzling anyway.

Here, my friend would not have gone joy-guzzling if she had not hot-wired her car. In turn, she would not have hot-wired her car if I had not hidden her keys. So, there is a sequence of counterfactual dependencies leading from my hiding her car keys to her joy-guzzling. According to SIMPLE WITH TRANSITIVITY, then, I caused her to go joy-guzzling by hiding her car keys. This might already make us think that there is something problematic about Lewis' early account of causation,<sup>11</sup> but more problems are looming on the horizon. If we agree that joy-guzzling might make a difference to which harms occur (as Eriksson does), my hiding my friend's car keys caused not only her going joy-guzzling, but also climate change and its related harms. Therefore, Lewis' early account together with THE HARM PRINCIPLE entail that I had a reason to refrain from hiding her car keys, because hiding them would cause climate-change-related harms. Surely, this is counterintuitive. If anything, I had a climate-change-related reason to hide her car keys.

You might think it is objectionable to treat another agent (my friend, in the example) in this objective way. However, the same problem would have arisen even if I had interacted with my friend in a more engaging way. Suppose that, instead of hiding her car keys, I try to the best of my ability to persuade her to come along for a bike ride instead, and that she kindly but firmly refuses and goes leisure driving instead. Again, in this example, SIMPLE WITH TRANSITIVITY entails that I caused her to go leisure driving, and, by extension, caused climate change and its related harms. If I had not tried to persuade her, she would not have refused to come along for the bike ride, and if she had not refused to come along for the ride, she would not have gone leisure driving. So, once again, SIMPLE WITH TRANSITIVITY combined with THE HARM PRINCIPLE wrongly entails that I had a climate-change-related reason *not* to try to persuade my friend to come along for a bike ride.

As these examples make clear, the idea that causation is transitive is at odds with the idea that a morally relevant cause contributes to its outcome.

---

<sup>10</sup> Here, I understand a counteractor of an outcome simply as the opposite of a contributor to an outcome. Similarly, I take counteractions to be negative contributions.

<sup>11</sup> Hall (2004) and McDermott (1995) would agree.

## Problems with Fragility

The fragility strategy runs into the same problem as the transitivity strategy, but for a different reason. Eriksson invokes the fragility strategy in order to show that a single drive (or a single flight) might cause climate-change-related harms. He argues for instance that if we describe the drought that actually occurred in its most minute details, we find that this drought would not have occurred if I had refrained from going for my drive, while some other, very similar, drought would have occurred instead. In that case, my drive would have caused the drought that actually ensued (described in every detail). Further, since I have reasons not to cause harm (as stated by THE HARM PRINCIPLE), I had a climate-change-related reason not to go for my drive.

So far so good. But consider this example: You set up and run a device that captures carbon dioxide from the atmosphere on a large scale and stores it in the bedrock. Did you cause climate change and its related harms by doing so? According to SIMPLE WITH TRANSITIVITY, you might have done so. To the extent that using this device made a difference to what harms occurred (described in every detail), you caused these harms. If you had refrained from using the device, the harms that actually ensued might not have occurred, but some other very similar harms might have occurred instead. So, by using your carbon capture device, you might have caused the climate-change-related harms that actually ensued (described in every detail). Further, since you have reasons not to harm others (according to THE HARM PRINCIPLE), it transpires that you had climate-change-related reasons not to use your carbon capturing device. This verdict must be mistaken. Again, we find that SIMPLE WITH TRANSITIVITY gives counterintuitive results because it does not distinguish between contributors and counteractors.

This problem arises in connection with SIMPLE as well. If we consider climate change and its related harms to be very fragile events, using a carbon capturing device might make a counterfactual difference to whether these harms occur.

Does this mean that SIMPLE and SIMPLE WITH TRANSITIVITY offer inaccurate pictures of causation? Not necessarily. Lewis accepts that counteractors might be causes. He asks us to consider the following case:

[COUNTERMOVE:] Imagine a conflict between Black and Red. [...] Black makes a move that, if not countered, would have advanced his cause. Red responds with an effective countermove, which gives Red the victory.

(Lewis 2000: 194)

About this case, he writes:



In many of these cases Red's victory would have come sooner, or more directly, without Black's move. Black's move prevents Red's victory as well as causing it: it causes one version, but it prevents another. If we thought we had to choose, we would wrongly infer that since it is a preventer it cannot be a cause.

(Lewis 2000: 194)<sup>12</sup>

It does not matter that Black's move counteracted Red's Victory. Black's move still caused Red's victory (described in detail). Counteractors can also be causes. Similarly, we might argue, it does not matter that you counteracted climate change by capturing carbon from the atmosphere. What you did caused the climate-change-related harms that actually ensued (described in detail).

Lewis might be right that there is some important causal notion that both counteractions and contributions satisfy. However, this notion is not relevant to THE HARM PRINCIPLE, and more generally, it is not relevant to decisions about the outcome-related reasons we have. The notion of causation we are looking for must be capable of distinguishing counteractions from contributions. The problem here is reminiscent of Kutz's complaint about Bunzl's account.

### **Can We Distinguish Contributions from Counteractions?**

Is it possible to resolve these problems by adding to SIMPLE WITH TRANSITIVITY an analysis of when one event contributes to an outcome, as opposed to counteracting it, and saying that we have a reason to refrain from performing an action if it contributes to harm? When it comes to climate change, it seems straightforward enough to say that emitting greenhouse gases into the atmosphere contributes to global warming (and future climate-change-related harms) and that removing the gases counteracts climate change (and the future harms). As Eriksson writes:

[E]mitting GHGs [i.e. greenhouse gases] rather than not emitting, emitting more rather than less, and emitting sooner rather than later, overall appear to increase the probability that one's actions cause harm.

(Eriksson 2019: 66)

More generally, it seems that adding to the underlying dimension in collective impact cases contributes to the outcome, and that subtracting from this dimension counteracts the outcome. Flipping a switch in HARMLESS TORTURERS contributes to the victim's pain, adding a drop of water in DROPS OF WATER contributes to the

---

<sup>12</sup> Lewis makes this point in relation to his later account of causation, but it can as well be made about his earlier account.

men's alleviation of thirst, voting for a certain party in a national election contributes to this party winning the election, and so on.<sup>13</sup>

However, once we have an account of contribution that can explain these intuitions, we can do without Lewis' account of causation.<sup>14</sup> It suffices to use our account of what it is to contribute to global warming, couple it with THE HARM PRINCIPLE, and obtain the result that I have a reason to refrain from going for a leisure drive because doing so contributes to climate change. There would be no need to appeal to SIMPLE WITH TRANSITIVITY to show this.

## Wright's NESS Account

The discussion of SIMPLE and SIMPLE WITH TRANSITIVITY has so far circled around pre-emption cases of different varieties. How about cases of (simultaneous) overdetermination such as ASSASSINS and THE LAKE? Lewis (1973a) sets such cases aside, saying that he doubts "common-sense opinions about them would be firm and uncontroversial enough to afford useful tests of the analysis" (194). Eriksson (2019) goes one step further and argues that "overdetermination seems impossible in the actual world", and that "alleged cases thereof will most likely collapse into instances of preemption and joint causation" (37).<sup>15</sup> Wright (1985, 1988, 2013) does not agree. As he sees it, cases of overdetermination are possible, and our intuitions about these cases are clear enough. For instance, he criticises SIMPLE by considering this case of what he takes to be overdetermination:

[BURNING DOWN THE HOUSE:] *C* and *D* independently start separate fires, each of which would have been sufficient to destroy *P*'s house. The fires converge and together burn down the house.

(Wright 1985: 1775-76)

---

<sup>13</sup> I think these intuitions are clear enough. However, if you understand contributions as making a marginal counterfactual difference to the outcome, you might not agree that voting for a certain party in a national election contributes to this party winning the election. More generally, you might not agree that acting in the relevant way contributes to the outcome in cases of overdetermination. That is okay. However, in that case, you do not have in mind the notion of contribution relevant to decisions about what outcome-related reasons we have. At least, so I will argue in the chapters following this one.

<sup>14</sup> My own preferred view is that contributions make the outcome more secure while counteractions make it less secure – an idea inspired by Touborg's (2017, 2018) account of causation. This account is explained in detail in Chapter 5 "Reasons for Action" and Chapter 11 "You Just Didn't Care Enough".

<sup>15</sup> Eriksson cites Bunzl (1979) and Paul and Hall (2013) on this issue.

Wright argues that each fire was a duplicative cause of the destruction of *P*'s house. Still, he continues, this is not the result SIMPLE gives. SIMPLE entails that neither *C*'s nor *D*'s fire was a cause of the destruction of *P*'s house. The same point can be made about SIMPLE WITH TRANSITIVITY. There is no intermediate event that both counterfactually depends on *C*'s starting a fire and is necessary for the destruction of the house. While *C*'s starting the fire was necessary for the fires to converge, the convergence of the fires was not necessary for the destruction of *P*'s house. *D*'s starting his fire guaranteed that the destruction would occur whether or not the fires converged. So, *C*'s starting a fire did not cause the destruction of the house according to SIMPLE WITH TRANSITIVITY. The same goes for *D*'s starting a fire. In general, in overdetermination cases, we will not find a cause by appealing to the transitivity of counterfactual dependence. In this respect, overdetermination cases and late pre-emption cases are alike.

In this chapter, I will dodge the question whether genuine overdetermination exists, and whether our intuitions about causation in these cases are sufficiently clear. I will proceed as if we can set aside cases of overdetermination when discussing SIMPLE and SIMPLE WITH TRANSITIVITY (as I already did); and as if our intuitions about such cases are clear enough to use them when theorising causation when discussing NESS. I will do this in hope of providing a charitable reading of each of these accounts. The complaints I have do not rely on our intuitions about overdetermination, so proceeding in this way should not invalidate the discussion. I merely use overdetermination cases to illustrate NESS.

Wright's (1985, 1988, 2013) account of causation delivers the verdict we are after in cases of overdetermination (i.e. that the overdetermining causes are indeed causes). On this account, a cause is a necessary element of a sufficient set. More precisely:

[NESS:] A condition *c* was a cause of a consequence *e* if and only if it was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the occurrence of *e*.

(Wright 2013: 18)<sup>16</sup>

It might be a good idea to tease out the various conditions incorporated in this formulation. NESS states that *c* was a cause of *e* just when there was a set of sufficient conditions for *e* such that: (1) *c* was a member of the set; (2) all elements

---

<sup>16</sup> NESS is similar to Mackie's (1965, 1974) INUS (Insufficient but Necessary element of a Unnecessary but Sufficient set) account of causation, and is inspired by Hume (1999/1748, 2007/1738-40), Mill (2011/1843) and Hart & Honoré (1985). For a discussion of the similarities and differences between NESS and INUS, see Wright (2013).

of the set obtained; (3) they obtained before *e* occurred; and (4) *c* was necessary for the sufficiency of the set.<sup>17</sup>

In BURNING DOWN THE HOUSE, two sets of existing antecedent conditions were sufficient for the occurrence of the destruction of *P*'s house. One consisted of *C*'s starting a fire, the current wind conditions being in a particular way, the presence of combustible material and oxygen, and so on. The other set consisted of *D*'s starting a fire, the current wind conditions being in a particular way, the presence of combustible material and oxygen, and so on. Now, unless *C* had started a fire, the first set would no longer have been sufficient for the destruction of *P*'s house. So, *C*'s starting a fire was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the destruction of *P*'s house. Teasing this out, we might say there was a set of conditions sufficient for the destruction of *P*'s house such that (1) *C*'s starting a fire belonged to this set, (2) all elements of the set obtained, (3) they did so at some time before *P*'s house was destroyed, and (4) this set would not have been sufficient for the destruction of *P*'s house had *C* not started a fire. Therefore, *C*'s starting a fire was a cause of the destruction of *P*'s house, according to NESS. Exactly the same goes for *D*.

Similarly, in the NESS framework, you, Vanessa and Walter each caused the collapse of the ecosystem in THE LAKE (see p. 66). The ecosystem was going to collapse if at least two of you used the hazardous paint. As things turned out, all three of you did. Even though none of you caused the collapse of the ecosystem according to SIMPLE (since this disastrous event would have occurred even if one of you had refrained from using the hazardous paint), what each of you did was a necessary element of a set of existing antecedent conditions that was sufficient for the collapse to occur. For instance, if you had used the environmentally friendly paint, the set consisting of you and Vanessa using the hazardous paint would no longer have been sufficient for the ecosystem to collapse. Overdetermined voting is another paradigmatic example of this kind, and so is ASSASSINS.<sup>18</sup>

---

<sup>17</sup> This reminds us of the way Wright (1988) defines NESS. I have added (3) to his definition.

<sup>18</sup> The examples given here are threshold cases. Does NESS fail to give the right verdict on causation in non-threshold cases? Consider climate change. NESS will only entail that the emissions from a single drive cause climate-change-related harm if these emissions are necessary for the sufficiency of an antecedent set that is sufficient for this harm to occur. In other words, NESS only entails that a single drive causes climate-change-related harm if there is a threshold at which this harm occurs. You might think there is no such threshold. Just as removing one grain of sand from a heap cannot turn a heap into a non-heap, a single drive could never determine whether or not some climate-change-related harm occurs. (I proposed this argument in Gunnemyr, 2019). However, as Touborg and I argue in Chapter 5 "Reasons for Action", this argument mistakenly presumes that there is this one event – some particular climate-change-related harm – with imprecise conditions of occurrence. Instead, we should take climate-change-related harms (and indeed any seemingly imprecise event) to consist of a range of different events, each with its own precise conditions of occurrence. With this more accurate view of vagueness, NESS will entail that a single drive causes climate change and its related harms.

Does NESS together with THE HARM PRINCIPLE explain why *C* and *D* had a reason not to start a fire, and why you had a reason to use the environmentally friendly paint in THE LAKE? It seems so. At least, they give the right verdict in these cases.

Does NESS describe the morally relevant causal relation we are looking for? Braham and Van Hees (2012) suggest that it does. Admittedly, they do not say much about outcome-related *reasons*, but instead suggest that NESS describes the causal relation that is relevant for *moral responsibility* – for blameworthiness and praiseworthiness. On their account, NESS-causation of an outcome is one of three independently necessary and jointly sufficient conditions of moral responsibility for this outcome. In addition, you must satisfy an agency condition (you did what you did intentionally), and an avoidance opportunity condition (you had a reasonable opportunity not to be morally responsible for this outcome). They show that this account gives intuitively correct verdicts on moral responsibility across a range of cases, including overdetermination cases and Frankfurt-style cases.

## Switching Cases

In other kinds of cases, however, NESS seems to deliver counterintuitive verdicts on moral responsibility and outcome-related reasons. Consider, for instance, the following switching case (which was introduced in Chapter 1):

THE ENGINEER: an engineer is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch, so that the train travels down the right-hand track, instead of the left. Since the tracks reconverge up ahead, the train arrives at its destination all the same; let us further suppose that the time and manner of its arrival are exactly as they would have been, had she not flipped the switch.

(Hall 2000: 205)

Let us add to this that the train arrived late at the station: it would have arrived late in the absence of the switch flipping and it did arrive late. Did the engineer's flipping her switch cause it to be late? And did she have a train-arriving-at-the-station-in-time-related reason not to flip the switch? It seems not. Her flipping the switch did not make any relevant difference to when or whether the train arrived at the station. However, this is not the verdict NESS and THE HARM PRINCIPLE give. The engineer's flipping of the switch was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the train to arrive late at the station,<sup>19</sup> a set that includes her flipping the switch, there being a right-hand track, a

---

<sup>19</sup> This is true whether we take sufficiency to be lawful sufficiency or causal sufficiency. (This distinction will be explained in Chapter 12, p. 254ff. I mention it here because Wright (2013) makes quite clear that he takes sufficiency to be causal sufficiency.) The engineer's flipping of the switch was necessary for the sufficiency of a set of existing antecedent conditions that

train that approaches in the distance, and so on. Without the flipping of the switch, this set of antecedent conditions would not have been sufficient for the train to arrive late at the station. So, the engineer's flipping of the switch was a cause of the train's arriving late at the station. This is counterintuitive. Worse still, when combined with THE HARM PRINCIPLE, NESS entails that the engineer had a reason not to flip the switch. Doing that caused the train to arrive late, which constitutes a harm. Just as Lewis' early account did, NESS entails that an event *c* might cause another event *e* even though *c* does not contribute to *e*.

Braham and Van Hees (2012) do not consider switching cases, but we can briefly note that this kind of case poses a problem, also, for their account of moral responsibility. While their account entails that the engineer is not blameworthy for the train's late arrival, they cannot explain, in the right way, why this is the case. On their account, the reason the engineer is not blameworthy for the train's arriving late is that she satisfies neither the avoidance condition nor the agency condition. Whether or not she flipped the switch, she would cause the train's late arrival at the station, and therefore she cannot avoid being blameworthy for this outcome. In addition, we might assume that she did not flip the switch with the intention of making the train arrive late at the station. However, these explanations seem to be of the wrong kind. Moreover, even if we accept them, there is also another explanation that we would like to be able to give. Surely, we want to be able to say that the engineer is not morally responsible for the train's arriving late because her flipping the switch did not cause the train to arrive late at the station. To be morally responsible for an outcome, you must contribute to it, and here what the engineer did (flipping the switch) did not contribute to the outcome. Therefore, she is not blameworthy for the train's late arrival. Since Braham and van Hees take NESS to be the relevant account of causation, they cannot give this explanation.

## Causes and Background Conditions

NESS creates at least two further problems. First, it gives counterintuitive verdicts in some late pre-emption cases. At least, it does so on a standard understanding of what it is for a set to be sufficient for an outcome. I will return to this issue in Part Two (on p. 254ff). Second, as I shall now explain using Schaffer's (2000) "Queen

---

*guaranteed* the train's arrival at the station, and it was a part of the set of conditions that was necessary for the outcome to occur *immediately* when each of these conditions had been satisfied, at least as long as we include the flipping of this switch among these conditions. I think we should allow that inclusion. If we do not, we will get into trouble in a case which is just like THE ENGINEER, but where the left-hand track is severed and the train will derail if it travels down that track. Here, we would want to say that the engineer caused the train's arrival at the station by flipping the switch. However, if we do not include the engineer's flipping of the switch in the relevant causal law, we cannot say this. (I see no reason why the relevant causal law should differ between the cases.) These remarks will become clearer in the later discussion of lawful sufficiency and causal sufficiency.

of England problem”,<sup>20</sup> it fails to distinguish causes from background conditions. Consider the following scenario:

WITHERING PLANT: Bob promised to water Joan’s plant while she was away. For some reason however, he failed to do so. As a result, the plant withered and died before Joan came back.

Here, it seems that Bob had a reason to water Joan’s plant, and that his omission to do so was a cause of the plant’s withering away. This is also the verdict that NESS, coupled with THE HARM PRINCIPLE, gives. Bob’s omission caused the plant’s withering away because it was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the plant’s withering away. Had he watered the plant, that set would no longer have been sufficient for the death of the plant. Further, Bob had a reason not to omit watering Joan’s plant because the omission caused harm. So far so good. The problem is that this reasoning applies to any agent who omitted to water Joan’s plant. For instance, the Queen of England caused the death of Joan’s plant, and therefore had a reason not to omit watering the plant. This is counterintuitive.

Braham and Van Hees (2012) suggest a promising solution to this problem. They follow John L. Mackie (1965) in arguing that causal evaluations are made in relation to particular causal fields. Ordinarily, we do not state what the relevant causal field is. Rather, it is given by the context. On this view, causation is not a two-placed relation between cause and effect, but a three-placed relation between cause, effect and causal field. As Mackie puts it:

A statement [... such] as “*A* caused *P*” is usually elliptical, and is to be expanded into “*A* caused *P* in relation to field *F*.”

(Mackie 1965: 249)

Equipped with the idea that causality is a three-placed relation, we can explain why the Queen of England did not cause Joan’s plant to wither away, while Bob did. The Queen’s omission is not part of the relevant causal field. What she did, or did not do, is simply irrelevant to the inquiry at hand. That said, it is not necessarily irrelevant. Her omission would become relevant if, for instance, she had also promised to water Joan’s plant. In everyday causal inquiries, however, where there is no reason to include the omissions, or inactions, of distant others, we simply do not treat it as a live possibility that the Queen of England would have watered Joan’s plant. The idea that a causal evaluation is relative to a certain causal field (or

---

<sup>20</sup> Hart and Honoré (1985), Woodward (2003); Beebe (2004), Sartorio (2004), McGrath (2005), Petersson (2019) and others discuss similar cases.

“possibility horizon”, which is the term Touborg and I use) will play an important role in the account of outcome-related reasons presented in the next chapter.

There are other solutions to the Queen of England problem. For instance, Sarah McGrath (2005) argues that causation has a normative component. The reason why Bob’s and not the Queen’s omission caused the plant’s withering away is that Bob but not the Queen had promised Joan to water the plant. Others, like James Woodward (2003), are prepared to accept that the Queen’s omission also caused the plant’s withering away, and suggest that it is for pragmatic reasons that we are reluctant to say that this is the case. The idea is the following. If you say that the Queen’s omission caused the plant’s withering away, others might easily read into this that it was “a serious possibility” (227) that the Queen had watered the plant. Since this further implicit implication is mistaken, we are reluctant to say that her omission caused the plant’s withering away.

My response to this issue (like Mackie’s) falls somewhere between these explanations. I agree with Woodward that we cannot say the Queen’s omission was a cause of the plant’s withering away unless there is “a serious possibility” that the Queen would have watered the plant, but I do not think that this is a merely pragmatic issue. Rather, I think there must have been a relevant possibility of the Queen watering the plant in order for her omission to be a cause of the plant’s withering away in the first place. I agree with McGrath that a further component guides judgements about causality, but I do not take this to be normative. I take it to concern which relevant possibilities there are. I will explain my response more carefully in the upcoming chapters.

## Conclusion

One way to reject the inefficacy argument is by responding with the claim that you might have an outcome-related reason to act in a certain way even when the relevant outcome will occur whether or not you act in this way. Various non-causal responses of this sort, including appeals to fairness, virtue ethics, and Kantian duties, are available. But as I argued in the previous chapter, these non-causal responses fail unless we can show that acting in the relevant way is a cause of the relevant outcome.

Quite a few writers have argued that you do not cause the outcome in collective impact cases. Acting in the relevant way (leisure driving, flipping a switch, etc.) is not a cause of the outcome, they typically argue, because the outcome would occur whether or not this act is performed. However, this argument is unsound. An act can be a cause of an outcome even if the outcome would occur whether or not this act is performed. The argument mistakenly relies on a simple counterfactual analysis of



causation. The problem is to find a more accurate analysis of causation that identifies the morally relevant causal connection we are after.

In this chapter, I have argued that neither Lewis' (1973a, 1986b, 1986a) early account of causation nor Wright's (1985, 1988, 2013) NESS condition is what we are looking for. Lewis' early account does not harmonise well with THE HARM PRINCIPLE. Since it takes causation to be transitive, and since it relies on the fragility strategy, it fails to distinguish between contributors and counteractors. It may capture an important causal notion (that you might cause an event even though you delay it, as in COUNTERMOVE). But it is clear that it does not capture the notion of causation relevant to outcome-related reasons. Wright's NESS account of causation does not fare much better. It gives counterintuitive verdicts in switching cases and fails to distinguish causes from background conditions. While the second of these problems can be solved by factoring in causal fields, the first looks more stubborn. In switching cases like THE ENGINEER, NESS entails that you cause the outcome even though you do not contribute to it.

Taking a step back and considering the broader picture, we see what seem to be two rather different causal notions: *contribution* and *causal involvement*. According to the first, a cause essentially contributes to its effect. According to the second, it does not. Contribution is relevant to what outcome-related reasons we have. Causal involvement is not. Lewis' early account is clearly an account of causal involvement (since counteractors can be causes, as illustrated by Lewis' discussion of COUNTERMOVE). NESS is also – perhaps less clearly – an account of causal involvement (since one event might cause another without contributing to it).<sup>21</sup>

What is it to *contribute* to an outcome? Here, as I did in the previous chapter, I have talked about contribution in a very loose sense, hoping that cases like CAR KEYS, HARMLESS TORTURERS and going for a leisure drive in a fossil fuel car give us a firm enough grasp of this notion. But here are some suggestions about what contribution might involve: an event E contributes to an outcome O when it increases the likelihood of O, makes O occur sooner rather than later, makes O occur rather than not occur, or makes O more secure (that is, O is further from not happening or closer to happening in case E occurs). There are probably other ways of understanding contribution as well.

Is causal involvement morally irrelevant? No, it is not. It is irrelevant to what outcome-related reasons we have, as cases like CAR KEYS and going for a leisure drive make clear. However, causal involvement is not irrelevant to moral responsibility. To be blameworthy (or praiseworthy) for an outcome, what you did must be part of the causal history of this outcome. This is particularly clear when

---

<sup>21</sup> Accounts of causation according to which a cause is part of the physical process that leads to an outcome also purport to describe causal involvement. Salmon (1994) and Dowe (2000) propose accounts of this sort.

we consider contribution, understood as probability-raising or security-raising. If you are to be blamed (or praised) for an outcome, you need to have done more than increase the likelihood of the outcome or make it more secure. You can increase the likelihood or security even when, as things turn out, the outcome does not occur, and obviously you cannot be blameworthy or praiseworthy for an outcome that never occurred. It appears to follow that, in order to be blameworthy or praiseworthy for an outcome, what you did must both contribute to the outcome and be involved in the causal history of it.<sup>22</sup>

These points are very handwavy. Still, they give a good idea of what is to come in the remainder of this thesis. Thus, (with Touborg) I will propose an account of outcome-related reasons. On this account, you have a reason to refrain from acting in a certain way if acting in that way would contribute to some harmful outcome. Contribution, here, is understood in terms of security-dependence within a possibility horizon (i.e. roughly, within a causal field). In Part Two, (again with Touborg) I will identify a pair of necessary conditions that are jointly sufficient for being blameworthy (or praiseworthy) for an outcome: very roughly, you are blameworthy for some outcome O just in case O is bad, you contributed to O and were involved in O's causal history. Somewhat more precisely, but still rather roughly, you are blameworthy for O just in case O is bad and you made O more secure and there is a causal process connecting you to O.

---

<sup>22</sup> This point is less clear if contribution is interpreted as making an outcome occur sooner rather than later, or as making an outcome occur rather than not occur. This is because, on these understandings, what you did is automatically part of the causal history of an outcome if you contributed to the outcome.



## 4. Non-Superfluous Causes

Before presenting my preferred account, I will consider what might be the most promising attempt to explain the reasons intuition so far, namely Nefsky's. She suggests that you have a reason to perform some action  $\varphi$  if doing so could help to bring about some beneficial outcome, and that you have a reason not to  $\varphi$  if  $\varphi$ -ing could help to bring about some harmful outcome. On her account, an action  $\varphi$  *could* help to bring about some outcome if it could be a non-superfluous part of the cause of this outcome.

I will first present Nefsky's account in further detail. I will then argue that if we pay closer attention to what it might mean for an act to be part of the cause of an outcome  $O$ , it might turn out that the additional conditions for non-superfluity are redundant. Finally, I will argue that Nefsky's account faces some important counterexamples. The critique I shall present will serve to motivate the positive account of outcome-related reasons given in the next chapter.

### Nefsky on Helping

Nefsky defines helping (and non-superfluity) as follows:

[HELPING:] Suppose your act of  $X$ -ing could be part of what causes outcome  $Y$ .

In that case, your act of  $X$ -ing is non-superfluous and so could help to bring about  $Y$  *if and only if*, at the time at which you  $X$ ,

(\*) It is possible that  $Y$  will fail to come about due, at least in part, to a lack of  $X$ -ing.

(Nefsky 2017: 2753)<sup>1</sup>

As I have done so far in this thesis, I will denote the relevant action  $\varphi$  and the collective outcome  $O$  in what follows, instead of using  $X$  and  $Y$  as Nefsky does.

For clarity, this definition needs unpacking. First, we have the supposition that (a) your act of  $\varphi$ -ing could be a part of what causes outcome  $O$ . Since your act of  $\varphi$ -ing

---

<sup>1</sup> Spiekermann (2014) has a similar suggestion.

will be a part of what causes outcome O only if O occurs, (a) implies that (b) it is possible that O will occur. Furthermore, the requirement that (c) it is possible that O will fail to occur, is implied by the requirement that (d) it is possible that O will fail to occur due, at least in part, to a lack of  $\varphi$ -ing. To bring out these implications, Nefsky's definition can be restated as follows:

HELPING: Your act of  $\varphi$ -ing could help to bring about outcome O if and only if at the time at which you  $\varphi$ ,

- (a) your act of  $\varphi$ -ing could be part of what causes outcome O,
- (b) it is possible that O will occur,
- (c) it is possible that O will fail to occur, and
- (d) it is possible that O will fail to occur due, at least in part, to a lack of  $\varphi$ -ing.

When these conditions are satisfied, and so when  $\varphi$ -ing helps to bring about O, Nefsky claims that you have a reason to  $\varphi$  if O is some beneficial outcome, and that you have a reason to refrain from  $\varphi$ -ing if O is some harmful outcome. I will call reasons of this sort *helping reasons*.

To illustrate Nefsky's account, let us again consider DROPS OF WATER. In this case, HELPING entails that pouring your pint into the water cart could help to alleviate the suffering of the people in the desert just in case the following conditions are satisfied:

- (a) pouring your pint into the water cart could be part of what causes the alleviation of suffering,
- (b) it is possible that the suffering will be alleviated,
- (c) it is possible that the suffering will not be alleviated, and
- (d) it is possible that the suffering will not be alleviated due, at least in part, to a lack of pints poured into the water cart.

If we grant that these conditions are satisfied, Nefsky's proposal yields the desired result: pouring your pint into the cart could help to bring about the alleviation of harm, and you have a helping reason to pour your pint into the cart.

Nefsky's proposal also captures the intuition that there are circumstances where you do *not* have a helping reason to pour your pint into the water cart. Suppose, for example, that the cart is already full at the time when you are contemplating pouring in your pint. In that case, condition (c) would not be satisfied: the good outcome would already be guaranteed, and you would therefore not have a helping reason to

pour in your pint. Pouring your pint into the cart could be part of the cause, but only superfluously so.

Two further clarificatory remarks are in place here. First, helping reasons are objective. Objective reasons are, roughly, the reasons we have when we have all of the relevant information about the situation. In contrast, subjective reasons are (roughly) the reasons we think we have as a result of the beliefs we actually have about the situation. These beliefs could be mistaken. If I did not know in DROPS OF WATER that the water in the cart was going to be distributed among the people suffering from thirst, I would be unaware that I have a reason to pour my pint into the cart. Still, objectively speaking, I would have such a reason.

Second, you can have helping reasons in a deterministic world. The possibilities that HELPING refers to are deliberatively relevant possibilities. Roughly speaking, they are reasons “we should regard as live possibilities in practical deliberative contexts” (Nefsky 2017: 2760). Nefsky does not fully explain how we should understand these possibilities. Instead, she allows readers to insert their own accounts of deliberatively relevant possibilities. She does, however, impose two constraints on what can count as live possibilities, one of which is that more than one possibility might be regarded as open even in a deterministic world. Possible worlds besides the actual world matter for what we have objective reasons to do.

## Paying Attention to Causation

Nefsky does not consider in detail what is required for your  $\phi$ -ing to be part of what causes an outcome O. Rather, she takes it to be an uncontroversial element in her description of the cases she considers that  $\phi$  could be part of the cause of O. For instance, she takes it to be uncontroversial that adding your pint of water in DROPS OF WATER could be part of what causes the men’s suffering to be alleviated:

*In Drops of water*, it is part of the set-up of the case that if enough of us donate our pints to the cart, this will result in the men’s suffering being alleviated. So, in that situation, individual acts of water donation will have been part of what caused the beneficial outcome. That’s not under debate.

(Nefsky 2017: 2750)

I disagree. Some serious metaphysical work is needed to get from the set-up of the case to the verdict that individual acts of water donation are part of what caused the outcome. Indeed, some widely used accounts of causation cannot deliver this result. To show that this is the case, I will consider again the simple counterfactual account of causation:

SIMPLE: an event  $C$  causes an event  $E$  if and only if: had  $C$  not occurred,  $E$  would not have occurred.

Since the alleviation of suffering never turns on a single individual act of pouring water into the cart, one such act does not cause the alleviation of suffering according to SIMPLE.

However, in her remarks above (and in HELPING) Nefsky does not say that individual donations of water might be causes of the beneficial outcome. She says they might be “part of the cause” of this outcome. This suggests that she does not have SIMPLE in mind, but rather something like the following:

SIMPLE-SET: An individual act  $\phi$  is part of what causes an event  $E$  if and only if:

- (i) if the set of events  $C$  had not existed,  $E$  would not have occurred, and
- (ii)  $\phi$  belongs to  $C$ .<sup>2</sup>

However, SIMPLE-SET does not straightforwardly provide the verdict Nefsky is seeking. It does not straightforwardly entail that pouring your pint into the cart could be part of what causes the alleviation of suffering. There are two problems here – one concerning premise (i) and the other concerning premise (ii). I will go through these in turn.

### **Does the Set Cause the Outcome?**

Consider the set of all the  $n$  acts of pouring water into the cart that are actually performed. Does this set of events cause the alleviation of the men’s suffering? According to SIMPLE-SET, the answer, surprisingly, is No. For this set to exist, all  $n$  acts of pouring water into the cart need to occur. If just one of the pourings in the set fails to occur, the set does not exist. The closest worlds where this set of events does not exist are worlds where just one of the  $n$  people fails to pour water into the cart, while the other  $n - 1$  people do. In these worlds, the men’s suffering is still alleviated. Thus, we get the verdict that the set of all  $n$  acts of pouring water into the cart does *not* cause the men’s suffering to be alleviated (see Petersson 2004, 2018).

For instance, say that 9,001 people pour their pints into the cart. Does the set of these 9,001 pourings cause the alleviation of the men’s suffering? No: the closest world(s) where this set of 9,001 pourings does not exist are worlds where 9,000 people pour their pints into the cart. In these, the alleviation of suffering is just the

---

<sup>2</sup> This idea can be traced back to at least Parfit (1984). Schwenkenbecher (2014) suggests something similar. See discussion on p. 62ff.

same, and therefore, SIMPLE-SET implies that the set consisting of 9,001 pourings does not cause the alleviation of the men's suffering.

You might think we should compare what happens in the actual world, not with what happens in the *closest* world(s) where the set does not exist, but rather with what happens when no one pours a pint into the cart. SIMPLE-SET allows for this. Although this world is further away from the actual one, it *is* a world where the set of 9,001 pourings does not exist. And if we take *this* world to be the relevant one for comparison, SIMPLE-SET yields the result that the set of 9,001 pourings causes the alleviation of harm. In the actual world, suffering is alleviated, but in the relevant counterfactual world where the set does not exist, suffering is not alleviated. However, even if we allow that the relevant comparison need not always be the closest world in which the set does not exist, there is a question why the relevant world would be the one in which no one pours a pint into the cart. Why do acts of pouring pints into the cart belong to the relevant set, and not, for instance, Fred Astaire's dancing in the distance? We need a principled way of deciding this. Otherwise, all we can say is that it is indeterminate whether the set of pourings caused the alleviation of suffering. This, I take it, is not enough to establish that you definitely have a helping reason to pour your pint into the cart.

### **Does the Act Belong to the Set?**

It seems that the issue whether SIMPLE-SET gives intuitively correct verdicts about causation boils down to the question of how we should decide which acts belong to the relevant set. How do we know when premise (ii) is satisfied, in other words? In Chapter 2, I considered some ways in which we might try to isolate the relevant acts (see p. 64ff). I considered whether the relevant acts are those that are performed with the intention of bringing about the outcome, those belonging to the smallest set of acts of which it is true that had none of these acts occurred then the outcome would not have occurred, or those that add to the underlying dimension of the collective outcome. I argued that these ways of isolating the relevant acts are either inaccurate or run into the superfluity problem.

Nefsky (2017, 2021) makes another suggestion. She appeals to *types* of act. To  $\phi$  is to perform an act of the relevant type. This is a natural suggestion, given her understanding of collective impact cases. According to her, these are cases where:

[t]here is some *type of act* such that if enough people act in that way, this—perhaps in combination with other things—will result in some morally significant outcome, and yet it seems that no individual act of that sort will itself make a difference with respect to this outcome.

(Nefsky 2017: 2745, my emphasis)



So, the idea is that pouring a pint into the cart in DROPS OF WATER could be part of what caused the alleviation of suffering, because it is true, of this type of act, that if enough people do it (i.e. pour their pints into the cart), suffering will be alleviated – yet no individual pouring makes a difference in suffering.

Still, when we appeal to types of act, act description matters a great deal. Consider the following variant of DROPS OF WATER:

GRAINS OF SAND. Everything is like in DROPS OF WATER, except that my pint is filled with sand instead of water. There are still 9,999 people around me, and their pints are filled with water. And there are still 10,000 people in the nearby desert suffering from thirst. If I pour my pint of sand into the cart, the sand will gather at the bottom, and when the water in the cart later is distributed among the people suffering from thirst in the desert, the sand will just remain at the bottom of the cart.

Intuitively, I have no reason to pour my pint of sand into the cart. Still, if we describe the relevant acts simply as being acts of pouring one's pint into the cart, my pouring of my pint into the cart belongs to the set that could bring about the alleviation of suffering. True, I would be pouring a pint of *sand*, but I would still pour my pint into the cart. This means, in turn, that all the conditions of HELPING are satisfied. (a) My act could be part of what causes the alleviation of suffering (since it is true that if enough people pour their pints into the cart, suffering will be alleviated); (b) it is possible that the suffering will be alleviated; (c) it is possible that it will not be alleviated; and (d) it is possible that it will be alleviated, at least in part due to a lack of people pouring their pints into the cart. So we obtain the result that, using the current act description, HELPING entails that pouring a pint of sand into the cart helps to bring about the alleviation of suffering – hence I have a reason to pour my pint of sand into the cart!

Have we used an inaccurate act description? The relevant acts are not simply acts of pouring one's pint into the cart, but acts of pouring one's pint of *water* into the cart. If we describe the relevant acts more specifically in this way, HELPING will give the intuitively correct verdict that pouring a pint of water into the cart helps to bring about the alleviation of suffering. It will no longer imply that I have a reason to pour my pint of sand into the cart. Condition (a) is no longer satisfied. My pouring my pint of sand into the cart does not belong to the relevant set of acts that could bring about the alleviation of suffering.

As can be seen, then, HELPING yields different verdicts depending on how we describe the relevant act in the case under review. Where such variation arises, Nefsky (2021) suggests that *we should always deploy the more specific act description*. This will always give the right verdict, she argues. This sub-rule works in GRAINS OF SAND, but it generates problems in other cases. Consider:

NON-STANDARD CAR. I have an old steam car, made at the beginning of the twentieth century. Like most cars, it runs on fossil fuel. In this respect, my car is perfectly standard. However, while most fossil fuel powered cars have an internal combustion engine, mine has a steam engine.<sup>3</sup>

The question is: Do I have a climate-change-related reason not to use my car? Intuitively, I do. Just as I have a climate-change-related reason not to use a standard fossil fuel car, I have such a reason not to use my steam car. There should be no morally relevant difference in this respect.

But what does HELPING say? If we describe the relevant type of act as using a fossil fuel powered car, we get the intuitively correct verdict that I have a climate-change-related helping reason not to use my steam car. (a) It could be part of the cause of climate change and its related harms, (b) it is possible that climate change and its related harms will occur, (c) it is possible that they will not occur, and (d), it is possible that they will not occur, at least in part due to a lack of people not using fossil fuel powered cars. However, if we describe the relevant type of act as using a fossil fuel powered steam engine car, we obtain a different verdict. My using my steam car would belong to a set that could never make any difference to the occurrence of climate change and its related harms. That is, (a) would not be satisfied. Moreover, it is not possible that these harms would not occur, at least in part due to a lack of people using fossil fuel powered steam cars. Hence, (d) is not satisfied either. Clearly, then, we arrive at one verdict about what reasons we have if we consider a more general act description, and another verdict about what reasons we have if we consider a more specific act description. Nefsky advises us to employ the more specific act description. However, in this case, that leads to the intuitively mistaken verdict that I lack a climate-change-related reason to refrain from using my fossil fuel steam car. Nefsky's advice seems unreliable.

I think there is a straightforward solution to this problem. Instead of saying that we should employ the more specific act description, Nefsky should have counselled us to opt for the relevant act description. In GRAINS OF SAND, the relevant act description is something like "pour your pint of water into the cart", or just "pour water into the cart". In NON-STANDARD CAR, the relevant act description is something like "use a fossil fuel powered car" or "use fossil fuels and release the emissions into the atmosphere".

The question now is whether there is any principled way to decide what the relevant act description is in any given case. I think there is. We can say that a relevant act description is any act description that accurately identifies acts contributing to the outcome. In GRAINS OF SAND, pouring a pint of water into the cart contributes to the

---

<sup>3</sup> A similar counterexample was suggested by Gunnar Björnsson during the Q&A after Nefsky's (2021) presentation.

alleviation of suffering, while pouring a pint of sand does not, so the relevant act description should be one that identifies acts of pouring pints of water into the cart (while excluding acts of pouring pints of sand into it). In NON-STANDARD CAR, driving a standard fossil fuel car contributes to climate change, and so does driving a fossil fuel powered steam car. So, the relevant act description in this case is one that includes “drives with a fossil fuel car”, but excludes, for instance, “drives with an electric car”. And so on. The problem is that if we accept this straightforward solution, the superfluity problem raises its head. Rather than providing a solution to the inefficacy problem, we must presuppose a prior solution to this problem. We must presuppose that there is a principled way of distinguishing contributions from non-contributions.

There could be some other way of distinguishing the relevant act description from the irrelevant alternatives. However, my suspicion is that we will have to choose between two evils. *Either* we shall have to say that a relevant act description is any act description that accurately isolates acts that contribute to the outcome. If we do that, we run into the superfluity problem – thus we will need an account of what it is to contribute to an outcome in order to assess act descriptions. *Or* we shall have to use some other way of assessing act descriptions. However, if we do that, we always run the risk of including irrelevant details (e.g. that my car has a steam engine) or overlooking relevant details (e.g. that I am pouring a pint *of sand* into the cart). That is, if we do not appeal to a prior account of what it is for an act to contribute to an outcome, we risk running into the disconnect problem.

This is perhaps enough to motivate a search for an account of what it is for an action to contribute to an outcome. However, for the moment I would like to consider another interpretation of Nefsky’s (2017) proposal.

## Physical Process

Nefsky (2017) may not be relying on the SIMPLE-SET account of causation. At several points, she seems to have in mind a production view of causation. Roughly speaking, views of this kind states that C is a cause of E just in case C is connected to E by a physical process consisting of atoms and other particles pushing each other or exerting forces on each other.

PHYSICAL PROCESS: C is a cause of E just in case C is connected to E by a physical process.<sup>4</sup>

---

<sup>4</sup> Production views of causation include the accounts suggested by Russell (1912-13), Aronson (1971), Salmon (1984), and Dowe (2000). For a different kind of production view, based on nomological connections rather than physical processes, see Hall (2004).

In her discussion of the following variant of DROPS OF WATER, for example, Nefsky seems to rely on a production account:

[POWER HOSE:] Suppose that ten thousand people each donate a pint and the cart is at maximum capacity. It is about to be driven into the desert, but then I come along, with my own supply of water – ten thousand pints worth. I take a power hose from my supply, lower it into the cart and turn it on full blast, forcing in the water from my supply.

(Nefsky 2017: 2751)

Commenting on this case, Nefsky writes:

my act is part of what in fact *causes* [the men's] suffering to be alleviated. After all, the majority of the water that the men actually drink is from my supply.

(Nefsky 2017: 2751)

This kind of reasoning is typical of a production view of causation.<sup>5</sup>

A production account of causation would entail that you have a helping reason to pour your pint of water into the cart in DROPS OF WATER. Since there is a physical process connecting each act of pouring water into the cart to the alleviation of the men's suffering, each such act causes the alleviation of suffering. However, production accounts like PHYSICAL PROCESS face other difficulties: they cannot accommodate omission-involving causation. Nefsky is aware of this problem, and suggests a solution – one that works in simple cases:

if acting in a certain way might make a non-superfluous causal contribution toward bringing about what is now an avoidable bad outcome, this can give one reason not to do so.

(Nefsky 2017: note 24)

In THE LAKE (presented on p. 66), for instance, you have a helping reason to refrain from using the cheaper, hazardous paint even though there is no physical process leading from that refraining to the survival of the fish in the lake. You have this reason because you could help bring about the death of the fish if you use the hazardous paint. A similar thing could be said about climate change: you have a

---

<sup>5</sup> Nefsky also uses a parking meter example (2017: 2752) to illustrate the point that we need to distinguish non-superfluous causes from superfluous ones. Again, when describing this example, she seems to have a production view in mind.

climate-change-related helping reason not to drive your fossil fuel powered car because you could help bring about climate change if you drive it.

In cases of so-called double prevention, however, this solution does not work. Consider the following variant of DROPS OF WATER:

GUTTER: There is one person in a nearby desert suffering from thirst. You are standing in front of a long gutter leading to this person. There is no one else around. You know that one pint of water will come flowing through the gutter soon. There is a removable obstacle in the gutter right in front of you. If the obstacle is removed, the person suffering from thirst will be able to collect the pint of water at the end of the gutter. If the obstacle is left in place, the water will not reach the person suffering from thirst. Instead, it will overflow, allowing you to collect it in an empty glass of yours.<sup>6</sup>

Intuitively, you have a reason to remove the obstacle because doing that could help the person suffering from thirst. However, this is not what HELPING entails if we assume a PHYSICAL PROCESS account of causation. In GUTTER, there is no physical connection between your removing the obstacle and the water flowing to the end of the gutter where the person suffering from thirst can collect it, so condition (a) of HELPING is not satisfied.

This point looks a little more persuasive when we focus on backward-looking moral responsibility (i.e. blameworthiness and praiseworthiness). Suppose we ask: Who is blameworthy for the person's continued suffering? Intuitively, you are blameworthy for this suffering if you do not remove the obstacle. However, if we require there to be a physical connection between what you do and an outcome for which you are blameworthy, we do not get this verdict. There is no physical connection between your omission to remove the obstacle and the person's not being relieved of their thirst.

Double-prevention cases abound (Schaffer 2000). For instance, when you turn on a fire hose in order to extinguish a fire, you remove an obstacle, letting the water spray on to the fire. When you perform the Heimlich manoeuvre on someone choking on food, you remove an obstacle, allowing this person to breathe again. If we take PHYSICAL PROCESS to be the relevant account of causation, HELPING will fail to deliver the intuitively correct verdict in all such cases.

At this point, I think the correct response is to say: so much the worse for the accounts of causation we have considered so far. My aim has merely been to show the following: to make Nefsky's account work, we need to pay careful attention to

---

<sup>6</sup> If necessary, this example could be scaled up to a non-threshold case in which ten thousand people are suffering from thirst (as in DROPS OF WATER) and ten thousand people could each contribute a pint of water by removing an obstacle.

the accompanying account of causation. When we do this, a further question arises: Might the right account of causation be able to do all the work that is needed, rendering Nefsky's conditions for being a *non-superfluous* part of a cause redundant? In "Reasons for Action" Touborg and I argue that this is indeed the case.

## Counterexamples to HELPING

HELPING faces other problems. I will here discuss two of them.

### Contrastive Reasons

Nefsky suggests that you have a helping reason to  $\varphi$  when  $\varphi$ -ing could be a non-superfluous cause of a beneficial outcome O. What, precisely, is it for O to be a *beneficial* outcome? Even if it is possible to categorise outcomes as beneficial or harmful in absolute terms, it does not seem plausible to say that this matters for our reasons for action. In absolute terms, for example, losing your leg is a clear enough case of a harmful outcome. However, if you are in a situation where the only alternative to losing your leg is losing your life, you will have reason to bring it about that you (merely) lose your leg. Likewise, the alleviation of harm does not necessarily involve causing a beneficial outcome in absolute terms. In DROPS OF WATER, for example, it only involves causing the men in the desert to suffer *less* than they did before. When we are concerned with reasons for action, then, I think the best interpretation of the conditions in which an outcome O counts as beneficial is comparative:

COMPARATIVE I: O is a beneficial outcome just in case there is at least one alternative outcome O\*, such that O and O\* are incompatible, and O is better than O\*.

With this interpretation, however, further difficulties arise. To see this, consider the following case:

TRAIN TRACKS: suppose you are standing by a switch for the railroad tracks. The switch has three settings: *express*, *local* and *broken*. If you set the switch to *express*, the train will arrive quickly at the station; if you set it to *local*, the train will arrive slowly at the station; and if you set it to *broken*, the train will derail. Suppose that the best of these three outcomes is that the train arrives quickly at the station; the second best is that the train arrives slowly at the station; and the worst is that the train derailed. Suppose further that the switch is initially set to *express*.

(Adapted from Schaffer 2012: 38)

Do you have a reason to move the switch from *express* to *local*? Intuitively, it seems that you do not. Doing so does not improve things. In fact, doing so makes things worse. However, on the current interpretation of what it is for an outcome to be beneficial, Nefsky's account delivers the result that you do have such a reason. Your moving the switch to *local* could (indeed, would) be a non-superfluous cause of the train's arriving slowly at the station. Furthermore, given COMPARATIVE I, the train's arriving slowly at the station is a beneficial outcome: it is incompatible with, and better than, the train's derailing. On Nefsky's account, it therefore seems that you have a helping reason to set the switch to *local*. (This is true also if we think of beneficial outcomes in absolute terms.) Of course, Nefsky's account also yields the result that you have a helping reason to leave the switch at *express*. But it remains a problematic implication of her account that you have a helping reason to move the switch from *express* to *local*.

Perhaps, the difficulty lies in the current suggestion about what it takes for O to be a *beneficial* outcome. With this thought in mind, we can see that the following suggestion fixes the above problem:

COMPARATIVE II: an outcome O is a beneficial outcome just in case, for every alternative outcome O\*, such that O and O\* are incompatible, O is better than O\*.

On this suggestion, however, Nefsky's account would fail to capture all the reasons you have. Consider, for example, an alteration of TRAIN TRACKS in which the switch is initially set to *broken*. In combination with COMPARATIVE II, Nefsky's account now yields the result that you do not have a helping reason to move the switch from *broken* to *local*, since the train's arriving slowly at the station is not a beneficial outcome. Intuitively, however, it seems that you *do* have a reason to move the switch from *broken* to *local*: after all, doing so averts disaster. It is true that you also have a reason to move the switch to *express*. Indeed – unless we add further details to the case – it seems that this is what you have most reason to do. However, it would seem false to say that you have no reason at all to move the switch from *broken* to *local*.<sup>7</sup>

Cases like TRAIN TRACKS thus present a problem for Nefsky's account. They indicate that reasons for action can be determined by contrasts: they can be concerned with whether your doing  $\varphi$  rather than  $\psi$  could be a cause of O rather than O\*, where  $\varphi$  and  $\psi$  are incompatible, O and O\* are incompatible, and O is better than O\*. Admittedly, HELPING could be revised to accommodate contrasts. We could revise it to say something along the following lines: Your act of  $\varphi$ -ing rather than  $\psi$ -ing is non-superfluous and so could help to bring about O rather than O\* if and only if, at the time at which you  $\varphi$  rather than  $\psi$ , it is possible that O rather than O\* will fail to come about due, at least in part, to a lack of  $\varphi$ -ing rather than  $\psi$ -

---

<sup>7</sup> Justin Snedegar (2017) gives a comprehensive defence of the idea that reasons are contrastive.

ing (supposing that  $\varphi$ -ing rather than  $\psi$ -ing could be part of what causes outcome O rather than O\*).

TRAIN TRACKS also reveals another problem. Nefsky’s account is insensitive to the current circumstances. Thus it is insensitive to the fact the switch is already set to *express* in the initial example and *broken* in the modified version of the example. It takes as input only what is possible, not what is actual. This problem, it seems to me, is not easily handled within the framework Nefsky works with. It also reappears in other types of case, including cases involving coordination problems.

## Coordination Problems

HELPING sometimes gives the wrong verdict in scenarios involving coordination problems, such as Hi-Lo games. Consider for instance the following case:

HI-LO GAME: We have decided to meet for lunch, but forgot to specify where. We always eat at one of two restaurants, one of which is closer for both of us, and we both know this. We can assume that closer is better, and that the restaurants are otherwise equally good. This is also something we both know.<sup>8</sup>

Here, there are four relevant possibilities, as depicted in the following figure:

		You...	
		...go to the closer restaurant	...go to the restaurant further away
I...	...go to the closer restaurant	Best	Worst
	...go to the restaurant further away	Worst	Second-best

Now, imagine that the following happens:

HI-LO GAME, CONTINUED: Today, you are deeply absorbed in thoughts about a particularly problematic philosophical problem on your way to our rendezvous, and suddenly you find yourself outside the restaurant further away. Also, for a similar reason, I am already there.

---

<sup>8</sup> A similar example is discussed by Kirk Ludwig (2017: 11).



On this day, it seems that I have no objective meeting-up-for-lunch-related reason to go to the closer restaurant. If I had known the relevant facts, I would have realised that we would not meet for lunch if I went to the closer restaurant. However, HELPING counterintuitively entails that going to the closer restaurant helps to bring it about that we meet up for lunch, and therefore it entails that I had an objective meeting-up-for-lunch-related helping reason to go to this restaurant. In the relevant circumstances it is true that: (a) my act of going to the closer restaurant could be part of what causes us to meet for lunch (in the eventuality that both of us go to this restaurant), (b) it is possible that we will meet up for lunch (in the eventuality that both of us go to the same restaurant), (c) it is possible that we fail to meet up for lunch (in the eventuality that we go to different restaurants), and (d) it is possible that we will fail to meet up for lunch due at least in part to a lack of visits to the closer restaurant (in the eventuality that that one of us, but not the other, goes to the closer restaurant). Again, we find that HELPING gives counterintuitive verdicts in a coordination problem.

We can note that, in a similar vein, HELPING entails that I have an objective minimise-my-sentence-related reason to stay silent in the prisoner's dilemma even if the other prisoner betrays me. I have this reason since there is a possibility that both of us will stay silent. Further, although I will not show this here, HELPING gives counterintuitive verdicts about cases like the Stag Hunt.<sup>9</sup> In "Reasons for Action", however, Touborg and I show that HELPING gives the wrong verdict in a pure coordination game where the agents' interests are aligned and all they have to do is to coordinate their choices. On our analysis, the reason why HELPING goes awry in all of these cases is that it is too focused on possibilities – it ignores which world is the actual one. Essentially, this is the same flaw that made HELPING give the wrong verdict about TRAIN TRACKS.

## Conclusion

The idea that acts can be *non-superfluous* parts of a cause relies on an inaccurate account of causation. It is not always clear what account of causation Nefsky has in mind, but it appears to be most likely that she is working with something like SIMPLE-SET, where a particular act is part of the set that could cause the outcome if it satisfies the appropriate act description. In some places, though, she seems to be employing a physical production account of causation. Each of these accounts is

---

<sup>9</sup> The example of the stag hunt was first introduced by Rousseau (1984/1755) as a prototypical example of the social contract. Hobbes (1997/1651) and Hume (2007/1738-40) have suggested other examples with a similar structure. Cases with this general structure are sometimes called "assurance problems". For more on this, see e.g. Skyrms (2004).

vulnerable to important counterexamples, and we should not be surprised that we encounter superfluous causes when using them.

HELPING also generates counterintuitive verdicts in cases like TRAIN TRACKS as a result of its insensitivity to contrasts. This problem can be avoided. Finally, I have argued that HELPING gives the wrong verdict in coordination problems and in cases like TRAIN TRACKS. Here, the root of the problem is HELPING's failure to take into account actuality – that is, which world the actual world is.

In the next chapter, Touborg and I suggest that if we simply pay closer attention to the metaphysics of causation, we can create an account explaining when you have outcome-related reasons on which an appeal to non-superfluity becomes redundant. This account avoids the superfluity problem and the disconnect problem. It also takes contrasts seriously and reflects the significance of the difference between possible worlds and the actual world.

Much of the discussion in Chapters 2 through 4 has been spelled out in terms of contributions. In the next chapter, Touborg and I give an account of outcome-related reason that takes seriously the idea that you have reasons to contribute to good outcomes, although we put it in terms of “promoting” rather than of “contributing”. We suggest, roughly, that an act promotes an outcome if and only if it makes the outcome more secure, and so that you have an outcome-related reason to perform an act if it makes some good outcome more secure.



# 5. Reasons for Action

Making a Difference to the Security of Outcomes<sup>1</sup>

(Paper 1)

**Mattias Gunnemyr and Caroline Touborg**

**Abstract.** In this paper, we present a new account of teleological reasons, i.e. reasons to perform a particular action because of the outcomes it promotes. Our account gives the intuitively right verdict in a number of difficult cases, including cases of overdetermination and non-threshold cases like Parfit's (1984) famous DROPS OF WATER. The key to our account is to look more closely at the metaphysics of causation. According to Touborg (2018), it is a necessary condition of causation that a cause increases the security of its effect. Building on this idea, we suggest, roughly, that you have a teleological reason to act in a certain way when doing so increases the security of some good outcome. This represents a middle way between the proposal that you have a reason to act in a certain way just in case this *would* cause a good outcome, and the proposal that you have a reason to act in a certain way just in case this *could* cause a good outcome.

---

<sup>1</sup> Unpublished manuscript.

# 1. Introduction

Many of the problems society faces today involve multiple agents: each individual agent makes no perceptible difference, but the result of thousands or millions of people acting in a particular way may nevertheless be catastrophic or save us all. Climate change is perhaps the most obvious example, but there are many more. The following case, originally proposed by Parfit (1984), brings out the crucial features and works as a metaphor for many other such cases:

[DROPS OF WATER:] Imagine that there are ten thousand men in the desert, suffering from intensely painful thirst. We are a group of ten thousand people near the desert, and each of us has a pint of water. We can't go into the desert ourselves, but what we can do is pour our pints into a water cart. The cart will be driven into the desert, and any water in it will be evenly distributed amongst the men.

If we pour in our pints, the men's suffering will be relieved. The problem is, though, that while together these acts would do a lot of good, it does not seem that any individual such act will make a difference. If one pours in one's pint, this will only enable each man to drink an extra *ten thousandth* of a pint of water. This is no more than a single drop, and a single drop more or less is too minuscule an amount to make any difference to how they feel.

(Nefsky 2017: 2743-44)

In cases like this, there is a strong intuition that each of us has a reason to donate our pint of water.<sup>1</sup> However, it has proved difficult to find a general account of reasons for action that supports this intuition. One of the most promising proposals is presented by Nefsky (2017). As Fanciullo (2020) argues, however, Nefsky's account faces serious counterexamples. In this paper, we therefore present a new account of reasons for action that supports the intuitive verdict on DROPS OF WATER.

More precisely, our account supports the claim that each of us has an *objective* and *pro tanto* reason to donate our pint. We focus on *objective* reasons for action in the sense that we defend the claim that each of us has a reason to donate our pint, even when all relevant information is taken into account. And we focus on *pro tanto* reasons in the sense that we defend the claim that we each have *a* reason to donate our pint. The claim that we each have a *pro tanto* reason is weaker than the claim that we each have *all-things-considered* reason to donate our pint, or that we are obligated to do so. However, it is sufficient to respond to the most obvious challenge to the intuitive verdict, namely that donating an extra pint makes no difference. As Nefsky writes:

---

<sup>1</sup> As Fanciullo (2020: 1488-89) points out, this intuition is supported by the thought that if everyone involved acts in accordance with their reasons, this will not lead to a morally suboptimal outcome. This thought is a version of the "principle of moral harmony", see Feldman (1980); Pinkert (2015).

When one says, “but it won’t make *any difference*”, more than just saying, “it doesn’t seem that I am obligated to act in that way”, one is saying “there doesn’t seem to be *any point at all* in acting in that way”.

(Nefsky 2017: 2744-45)

By showing that we each have *a* reason to donate our pint, we respond to the challenge: we show that there *is* a point in adding an extra pint.

To show this, we develop a unified account of *teleological* reasons for action. A teleological reason to  $\varphi$  is a reason to  $\varphi$  that is grounded in the fact that  $\varphi$ -ing *promotes* a certain outcome – for example, the outcome that the men’s suffering is relieved (see e.g. Portmore).<sup>2</sup> A crucial question in understanding when you have a teleological reason to  $\varphi$  is: what is the relevant relation of *promoting*?

In the following, we propose a new answer to this question. Roughly, we propose that *promoting* should be understood in terms of *making an outcome more secure* (see Section 4). Even if the men’s suffering is not in fact fully relieved, donating your pint makes this good outcome more secure: it brings us a step closer. This proposal is motivated by two considerations. First, it successfully handles cases that present problems for rival accounts. Second, it is theoretically motivated: it establishes a clear connection between causing and promoting, and upholds key inferences.

We proceed as follows. First, we set out three starting assumptions (Section 2). Second, we argue that an account of promoting needs to capture certain key inferences, and we present a number of test cases for our account (Section 3). Next, we set out our account, and show that it upholds the key inferences (Section 4), and we show that our account delivers intuitively correct verdicts on our test cases (section 5). Finally, we show that our account supports the intuitive verdict in DROPS OF WATER (Section 6).

## 2. Starting Assumptions

For the sake of simplicity, we assume in the following that the laws of nature are deterministic. Furthermore, we rely on three assumptions about teleological reasons.

First, we assume that the actions you have reason to do are time-indexed: you do not simply have a reason to  $\varphi$ . Rather, you have a reason to  $\varphi$  at time  $t$ . Or, in the

---

<sup>2</sup> Standardly, reasons are taken to be facts. For example, the fact that it is raining is a reason for you to bring an umbrella. This way of thinking about reasons may be connected to questions about promoting as follows: a fact  $F$  is a reason for you to  $\varphi$  just in case  $F$  explains why your  $\varphi$ -ing promotes some good outcome. For example, the fact that it is raining explains why bringing an umbrella promotes the outcome that you remain dry. See Schroeder (2007).

case of temporally extended actions or omissions you may have a reason to *begin* to  $\varphi$  at time  $t$ .<sup>3</sup>

Second, we assume, following Snedegar (2017), that teleological reasons are contrastive: you do not simply have a teleological reason to  $\varphi$  at  $t$ ; you have a teleological reason to  $\varphi$  rather than  $\psi$  at  $t$ , where  $\varphi$  and  $\psi$  are two incompatible actions or omissions in the sense that it is not possible for you to both  $\varphi$  and  $\psi$  at  $t$ . Furthermore, you do not merely have such a reason in virtue of how your action relates to some outcome  $O$ ; you have such a reason in virtue of how your action relates to whether outcome  $O$  will occur *rather than some incompatible outcome*  $O^*$ . Our assumption that teleological reasons are contrastive can be motivated by the very same cases that motivate a contrastive account of causation, such as:

TRAIN TRACKS: Suppose that you are standing by a switch in the railroad tracks. The switch has three settings: *express*, *local*, and *broken*. If the switch is set to *express*, the train will arrive quickly at the station; if the switch is set to *local*, the train will arrive slowly at the station; and if the switch is set to *broken*, the train will derail. Suppose that of these three outcomes, the best outcome is that the train arrives quickly at the station; the second-best is that the train arrives slowly at the station; and the worst outcome is that the train derailed. Suppose further that the switch is initially set to *broken*.

(Cf. Schaffer 2012: 38)

Some intuitions about this case may be difficult to capture without the resources of contrastivism. Consider the following two claims:

- (i) You have a reason to move the switch to *local* rather than moving it to *express*.
- (ii) You have a reason to move the switch to *local* rather than leaving it at *broken*.

Intuitively, we think that (i) is false, while (ii) is true. Since the only difference between these two claims lies in the choice of contrast, we need to go contrastive in order to accommodate both of these verdicts.<sup>4</sup> To avoid cumbersome repetitions, however, we will sometimes leave out the contrasts in what follows, saying simply that “you have a reason to  $\varphi$ ”, when it is obvious what the relevant contrast  $\psi$  is.

We acknowledge that contrastivism is not currently the standard view about reasons. It is worth noting, therefore, that our account could be made non-contrastive, though

---

<sup>3</sup> Skorupski (2010) argues that reasons are time-indexed. This raises a further question about *when* you have a reason to  $\varphi$  at  $t$ : do you only have this reason at  $t$ , or do you also have it some time prior to  $t$ , or even after  $t$ ? We stay neutral on this question.

<sup>4</sup> Sinnott-Armstrong (2008) argues in favour of contrastivism about reasons in a similar fashion. For further discussion, see Snedegar (2017: 25-44).

it would then need to be supplemented with some alternative way of handling cases like *Train Tracks*.

Third, we assume that you only have a reason to  $\varphi$  rather than  $\psi$  at time  $t$  when it is an option for you to  $\varphi$  at  $t$ , and also an option for you to  $\psi$  at  $t$ . Later on, we shall say more about how we understand the notion of an option. But for now, it is enough to note that we understand it in a natural, common-sense way, where it is often true that you have several different options open to you at a given time. In *Drops of water*, for example, you have at least two options: you have the option of donating your pint, and the option of keeping it to yourself.

Given these three assumptions, we may state our question more precisely. The question is how to fill in the blank in the schema below, in a way that captures how your  $\varphi$ -ing rather than  $\psi$ -ing promotes the occurrence of outcome O rather than O\*:

SCHEMA: you have a teleological reason to  $\varphi$  rather than  $\psi$  at time  $t$ , where  $\varphi$  and  $\psi$  are two mutually incompatible actions or omissions, just in case

- (a) it is an option for you to  $\varphi$  at  $t$ ,
  - (b) it is an option for you to  $\psi$  at  $t$ , and
- there are two incompatible outcomes O and O\*, such that
- (c) O is better than O\*, and
  - (d) [fill in the blank].

### 3. Desiderata

In this section, we consider three prominent suggestions about how to fill in the blank in the schema above:

WHETHER-WHETHER DEPENDENCE:

- (d) whether O or O\* occurs depends on whether you  $\varphi$  or  $\psi$  at time  $t$ .<sup>5</sup>

CAUSE:

- (d) your  $\varphi$ -ing rather than  $\psi$ -ing at time  $t$  would be a cause of O rather than O\*.

POTENTIAL CAUSE:

- (d) your  $\varphi$ -ing rather than  $\psi$ -ing at time  $t$  could be a cause of O rather than O\*.

---

<sup>5</sup> A further, natural suggestion is that promoting should be understood in terms of probability-raising (Schroeder 2007). In a deterministic setting, this suggestion is extensionally equivalent to WHETHER-WHETHER DEPENDENCE.



Consequentialists are typically committed to a non-contrastive version of WHETHER-WHETHER DEPENDENCE; Braham and van Hees (2012) defend a non-contrastive version of CAUSE; and Nefsky (2017) defends a non-contrastive version of POTENTIAL CAUSE. By understanding the advantages and disadvantages of these suggestions, we can get a better picture of the desiderata that a successful account of teleological reasons needs to satisfy.

Let us begin by considering WHETHER-WHETHER DEPENDENCE. We think this suggestion successfully captures a *sufficient* condition for when you have a teleological reason: whenever it is the case that O would occur if you were to  $\varphi$  at  $t$ , and O\* would occur if you were to  $\psi$  at  $t$ , you have a reason to  $\varphi$  rather than  $\psi$  at  $t$ . Holding on to this principle is a desideratum for any account of teleological reasons.

However, WHETHER-WHETHER DEPENDENCE fails to give a necessary condition for when you have a teleological reason. This becomes clear already in simple overdetermination cases like the following:

NUCLEAR SAFETY: you and Suzy work as engineers at a nuclear power plant. You independently notice that there is a problem. At time  $t$  you each press a button to safely shut down the reactor. Each button-pressing is an overdetermining cause of the shut-down of the reactor. If just one of you had pressed your button at time  $t$ , the reactor would still have shut down. But if neither of you had pressed your button at time  $t$ , there would have been a nuclear disaster.

It seems attractive to be able to say that the reactor was shut down safely, rather than melting down in a nuclear disaster, because you both acted in accordance with what you had objective reason to do. This supports the verdict that Suzy had a reason to press her button at time  $t$ , and that you did too.<sup>6</sup>

However, WHETHER-WHETHER DEPENDENCE delivers the result that *neither* you nor Suzy had a reason to press your buttons. Since Suzy pressed *her* button, it did not depend on your actions whether the reactor would be shut down safely or there would be a nuclear disaster. And since you pressed *your* button, it did not depend on Suzy's actions either.<sup>7</sup>

---

<sup>6</sup> If you do not share this verdict, it is probably because you do not take seriously the possibility that Suzy might not have pressed her button. If we hold fixed that Suzy presses her button, we agree that you had no reason to press yours. However, as we suggest later, we should generally take seriously the possibility that other agents might have acted differently. See also footnote 17, this chapter.

<sup>7</sup> Consequentialists like Singer (1980), Norcross (2005) and Kagan (2011) are in principle committed to WHETHER-WHETHER DEPENDENCE. Thus, they are committed to saying that you did not have an *objective* reason to press your button in NUCLEAR SAFETY. They mitigate the cost of this position by pointing out that you did have a *subjective* reason to press your button: assuming that you did

One alternative is to appeal to CAUSE instead of WHETHER-WHETHER DEPENDENCE (as e.g. Braham and van Hees do). Again, we think that CAUSE succeeds in capturing a *sufficient* condition for when you have a teleological reason: whenever it is the case that your  $\varphi$ -ing rather than  $\psi$ -ing at time  $t$  would be a cause of O rather than O\*, you have a reason to  $\varphi$  rather than  $\psi$  at time  $t$ . Holding on to this principle is once again a desideratum for an account of teleological reasons.<sup>8</sup>

CAUSE also delivers the intuitively correct result in NUCLEAR SAFETY, provided that it is combined with an account of causation that allows for overdetermining causes. If we count your button-pressing as an overdetermining cause of the safe shut-down, it immediately follows from CAUSE that you had a reason to press your button. Similarly, Suzy had a reason to press her button.

However, we do not have to look far to find cases that create trouble for CAUSE as well. Consider, for example, the case below:

THE LAKE: You, Vanessa and Walter all live close to a lake with a sensitive ecosystem. Each of you have a boat. If two or more of you paint the hull of your boat with a cheap and toxic antifouling rather than a more expensive but non-toxic one, the ecosystem in the lake will collapse. If at most one of you uses the toxic paint, the ecosystem will continue to thrive. As it turns out, all three of you use the cheaper paint, and the lake becomes a wet wasteland.

(Adapted from Björnsson 2014)

Did you, as an individual, have a reason to use the non-toxic paint rather than the toxic one? We think you did. Indeed, we think that all three of you had such a reason, and that the bad outcome happened because all three of you failed to act in accordance with this reason.

However, CAUSE yields the result that you, as an individual, did not have such a reason. Remember that CAUSE says, roughly, that you have a reason to use the non-toxic paint just in case your doing so *would* be a cause of life in the lake continuing to thrive. If you had used the non-toxic paint in THE LAKE, however, Vanessa and Walter would still have used the toxic paint, and so the ecosystem would still have collapsed. Your using the non-toxic paint therefore would not have been a cause of the ecosystem's continuing to thrive, for the simple reason that the ecosystem would

---

not know what Suzy would do, the expected utility of pressing your button was higher than the expected utility of not pressing your button, since pressing your button *might* make the difference between safely shutting down the reactor and a nuclear disaster.

<sup>8</sup> Nefsky (2017: 2757) uses a counterexample (the parking meter example) to argue that CAUSE does not give a sufficient condition for having a reason. In arriving at her verdict on what causes what in the counterexample, Nefsky relies on causal transitivity. However, as many have pointed out, transitivity fails precisely in cases with this structure. (See e.g. Paul and Hall, 2013: 232-244). We therefore believe her argument fails.

not have continued to thrive. Thus, CAUSE yields the verdict that you had no teleological reason to use the non-toxic paint in THE LAKE. (More precisely, you had no such reason related to whether the ecosystem would survive or collapse.) And the same applies to Vanessa and Walter. CAUSE therefore implies that none of you had a teleological reason to use the non-toxic paint rather than the toxic one.

Finally, let us consider POTENTIAL CAUSE, which is the least demanding of the three conditions. POTENTIAL CAUSE is satisfied in both of the cases we have considered so far: since your pressing your button *was* an overdetermining cause of the reactor shutting down safely, rather than melting in a nuclear disaster, it obviously *could* be. And likewise, your using the non-toxic paint *could* be a cause of the ecosystem thriving rather than collapsing, since it would be a cause if Walter or Vanessa had also used the non-toxic paint.

However, whereas WHETHER-WHETHER DEPENDENCE and CAUSE encounter problems because they are too demanding, POTENTIAL CAUSE runs into trouble because it is not demanding enough. To illustrate this, consider the following case:

COORDINATION: As part of a game-show, you and Sally are placed on opposite sides of a wall. Each of you is given a choice between either raising your hand or not at a given signal. If you both raise your hands, or both do not, you will receive a million dollars each. If one of you raises your hand, and the other does not, you will receive nothing. You raise your hand at the given signal, and Sally does too. You win a million dollars each.

In this case, it is clear that you had a reason to raise your hand when the signal was given (call this time  $t$ ) rather than keeping it down. It is also attractive to hold that you did *not* have any reason to keep your hand down at time  $t$ . Indeed, if you had kept your hand down at time  $t$ , you and Sally would both have received nothing.<sup>9</sup>

However, POTENTIAL CAUSE yields the verdict that you did have a reason to keep your hand down at time  $t$ : presumably, it is a relevant possibility that Sally might have kept her hand down at time  $t$ . And if Sally had kept her hand down, your keeping your hand down at time  $t$  would have caused each of you to receive a million dollars rather than nothing. Thus, keeping your hand down *could* have been a cause of the good outcome. In this case, we think that POTENTIAL CAUSE admits reasons that simply are not there.

---

<sup>9</sup> Note that our verdict on COORDINATION is consistent with taking seriously the possibility that Sally might have acted differently. As the case is construed, there are two possibilities in play: Sally could either raise her hand or keep it down. Still, only one of these possibilities occurs in the actual world (namely, that she raises her hand). POTENTIAL CAUSE yields a mistaken verdict exactly since it merely focuses on possibilities without taking into account which possibilities are actual.

Nefsky's account (2017) is a version of POTENTIAL CAUSE. According to Nefsky, you have a reason to  $\varphi$  just in case  $\varphi$ -ing *could* help, where helping consists in making a non-superfluous causal contribution. Not surprisingly, COORDINATION therefore presents a problem for Nefsky's account: in the possible world where neither you nor Sally raise your hands, keeping your hand down makes a non-superfluous causal contribution to your getting a million dollars each. Thus, keeping your hand down *could* help. And so, Nefsky's account yields the counterintuitive result that you have an objective reason to keep your hand down at  $t$ .

By considering these three suggestions, we now have clear desiderata for an account of teleological reasons. First, it needs to respect the sufficiency of WHETHER-WHETHER DEPENDENCE and CAUSE. Second, it needs to accommodate our intuitive verdicts in the three test cases, NUCLEAR SAFETY, THE LAKE, and COORDINATION. In order to achieve this, we need to find a condition that represents a middle way between CAUSE and POTENTIAL CAUSE, by being less demanding than CAUSE, but more demanding than POTENTIAL CAUSE. We set out how to do this in the following section.

## 4. Finding a Middle Way

To find a middle way between CAUSE and POTENTIAL CAUSE, we pay closer attention to the metaphysics of causation: POTENTIAL CAUSE is a weaker version of CAUSE because it merely requires that your  $\varphi$ -ing rather than  $\psi$ -ing at  $t$  *could* (rather than *would*) be a cause of O rather than O\*. However, once we pay attention to the metaphysics of causation, another attractive way of weakening CAUSE comes into view: if there are several necessary and jointly sufficient conditions for causation, we may weaken CAUSE by requiring only that some, but not all, of these conditions are satisfied. That is precisely the idea we develop in the following.

To do so, we begin from the account of causation developed by Touborg (2018). According to this account, there are two individually necessary and jointly sufficient conditions for causation: first, the condition of *process-connection*, which requires that a cause must be connected to its effect via a genuine process; second, the condition of *security-dependence within a possibility horizon*, which requires that a cause must make a difference to the security of its effect. We may weaken CAUSE by merely requiring that one of these two conditions is satisfied, rather than requiring the satisfaction of both. The condition of security-dependence within a possibility horizon is a highly promising candidate for this move.<sup>10</sup>

---

<sup>10</sup> While the idea for this move springs from the metaphysics of causation, our account of reasons is in principle independent of how the debate about causation turns out.

Consider again NUCLEAR SAFETY where you and Suzy both press your safety buttons, and where one such pressing is sufficient for safely shutting down the reactor. Even though neither you nor Suzy made a difference as to *whether* the outcome was going to occur, there is a sense in which each of you made the safe shutdown of the reactor *more secure*. As it happened, two things stood in the way of a nuclear disaster: your pressing your button, and Suzy’s pressing hers. If you had not pressed your button, only one thing would have stood in the way of a nuclear disaster: Suzy’s pressing her button. The same reasoning applies to Suzy. In this way, each of you increased the security of the safe shutdown of the reactor. We think it is precisely because of this that both of you had a reason to press your safety buttons: by pressing your button rather than not, you made the safe shutdown of the reactor *more secure*, while making a nuclear meltdown *less secure*.

To make this more precise, it is useful to think about security in terms of the distance-at-a-time between possible worlds. Considering two possible worlds  $w$  and  $w^*$ , let us say that  $w$  is close-to- $w^*$ -at-time- $t$  to the extent that the complete state of world  $w$  at  $t$  is similar to the complete state of world  $w^*$  at  $t$ . Based on this, we may give an initial definition of security as follows:

SECURITY: the security of outcome  $O$  in  $w$  at  $t$  is given as follows:

If  $O$  occurs in  $w$ , then  $O$  has positive security in  $w$ , and its degree of positive security in  $w$  at  $t$  is given by the distance-at- $t$  between  $w$  and the closest-to- $w$ -at- $t$  world(s) where  $O$  does not occur.

If  $O$  does not occur in  $w$ , then  $O$  has negative security in  $w$ , and its degree of negative security in  $w$  at  $t$  is given by the distance-at- $t$  between  $w$  and the closest-to- $w$ -at- $t$  world(s) where  $O$  occurs.

With this initial definition, our proposal is going to be, roughly, the following way of filling in the blank in the schema, using “@” to denote the actual world:

SECURITY-DEPENDENCE:

- (d)  $O$  is more *secure* and  $O^*$  is *less secure* at  $t$  in the closest-to-@-at- $t$  world(s) where you  $\varphi$  at  $t$  than they are in the closest-to-@-at- $t$  world(s) where you  $\psi$  at  $t$ .

To make this proposal fully precise, we need to answer the following question: which worlds should be taken into consideration when we are looking for “the closest-at- $t$ -worlds where ...”? Should we consider *all* possible worlds, in the widest sense, or should we make use of a restricted notion of possibility? Like Nefsky (2017), we think that the notion of possibility that is relevant here is a restricted notion of possibility. Nefsky suggests that the relevant notion of possibility is

the standard notion of possibility that we use in contexts of practical deliberation. That is, the possibilities that come into the account are those that we should regard as live possibilities in practical deliberative contexts.

(Nefsky 2017: 2760)

Nefsky continues: “It is debatable which notion of possibility we do use in such contexts and, for the purposes of this paper, I want to remain as open as possible about this.” (Nefsky 2017: 2761). So do we. However, we agree with Nefsky that the relevant notion of possibility needs to satisfy the following two constraints:

First, there is a distinction between “what is *possible* and what you have *reason to believe is possible*” (Nefsky 2017: 2761). These two may come apart: something may be possible, even though you have reason to believe it is not, and *vice versa*. According to Nefsky, objective reasons are based on what is *possible*; subjective reasons are based on what you have *reason to believe is possible* (Nefsky 2017: 2761). Since we are here concerned with objective reasons, we impose the following constraint: the relevant notion of possibility is concerned with what is *possible*, and what is possible may be different from what you have *reason to believe is possible*.

Nefsky’s second constraint is based on the observation that, in contexts of practical deliberation, “we think of agents (both ourselves and others) as typically being able to choose between several different courses of action, where different outcomes can result depending on what they choose to do” (Nefsky 2017: 2762). The relevant notion of possibility should capture this.

To do so, the relevant notion of possibility needs to include *enough* possibilities: when you deliberate about what to do, you take yourself as well as other agents involved in the situation to be able to choose between several different courses of action. When you deliberate about what to do in NUCLEAR SAFETY, for example, you take yourself to have the option of pressing the safety button, as well as the option of not pressing the button, and you take Suzy to have the same options. And when you deliberate about what to do in THE LAKE, you take yourself to have the option of using the non-toxic paint, as well as the option of using the toxic paint, and you take Vanessa and Walter to have the same options.<sup>11</sup>

At the same time, the relevant notion of possibility should not include *too much*: not just any possibility is a relevant – or “live” – option (see Nefsky 2017: 2763). In this way we avoid what Streumer (2007) calls “crazy reasons”: since doing so is not an

---

<sup>11</sup> This might seem to be a flat rejection of determinism. This is not what we have in mind (nor is it what Nefsky has in mind). The point is rather that we must take certain possibilities seriously *even if* it is determined by the past and the laws of nature that they will not be actualised: even if determinism is true, the relevant notion of possibility in deliberative contexts is such that we do and should think of ourselves and others as being able to choose between several incompatible courses of action.

option for you, you could never have reason to for instance go back in time and singlehandedly prevent slavery, the crusades and the two world wars.<sup>12</sup>

We may use the notion of a *deliberatively relevant possibility horizon* to capture this: a *possibility horizon* is simply a class of possible worlds, and the *deliberatively relevant possibility horizon* for the purpose of determining what you have reason to do at time  $t$  is the class of worlds that contains just those possible worlds that are relevant to determining what you have reason to do at  $t$ . We suggest the following procedure for arriving at the deliberatively relevant possibility horizon: Consider the agents involved in the situation in question. Each of these agents has certain actions and omissions that are open to her at time  $t$ . These are the agent's *options* at time  $t$ . A choice assignment is a specification of how each agent chooses among the options that are open to her at  $t$ .<sup>13</sup> In the case of NUCLEAR SAFETY, for example, there are four choice assignments, representing every combination of your choice {press, do not press} and Suzy's choice {press, do not press}. Every such choice assignment represents a deliberatively relevant possibility.<sup>14</sup> As a minimum, the *deliberatively relevant possibility horizon*  $H(t)$  for determining what each agent has reason to do at time  $t$  should include worlds representing every such choice assignment.<sup>15</sup> As a rule of thumb, no further possibilities need to be represented. However, this is merely a rule of thumb: although we do not consider any such cases here, there may well be cases in which further possibilities – possibilities that are not based on the options available to agents – should also be represented.<sup>16</sup>

---

<sup>12</sup> We agree with Streumer that “Reasons imply Can”. Note, however, that our notion of an option is different from his: unlike Streumer, we believe that determinism is compatible with the claim that you often have more than one option. For an overview of the discussion regarding “Reasons imply Can”, see Werkmäster (2019). For discussion, see e.g. Jeppsson (2016).

<sup>13</sup> In some cases, some combinations of options may be impossible. For instance, it is not an option for you to dance the tango with me if I'm not willing to dance the tango with you. In this paper, we set aside options that involve joint action, and focus on actions and omissions that are options for you no matter what others do. This means that, in the cases we discuss, every combination of the options we consider is possible.

<sup>14</sup> See Nefsky (2017: 2762, footnote 37).

<sup>15</sup> Of course, merely specifying e.g. what you and Suzy do – for example, that neither of you press your safety button – is not enough to fully characterize a possible world. We assume that the relevant world(s) representing this possibility start out by being as similar as possible (consistent with neither of you pressing your buttons) to the actual world at time  $t$ , and then evolve forwards from there in accordance with the actual laws of nature. Cf. the method for evaluating counterfactuals proposed in Paul and Hall (2013: 47-49).

<sup>16</sup> For example, Fanciullo (2020) considers a version of DROPS OF WATER where the 9999 other agents are replaced by mechanisms. He reports that “my intuition that you can help in 9999 Mechanisms is, admittedly, somewhat weaker than my intuition that you can help in Drops of Water” (1493). Our account can capture this. In 9999 Mechanisms, our rule of thumb does not settle what the relevant possibility horizon is. If it only includes the options that are open to *you*, you have no reason to add your water. However, it may also include further possibilities that are not based on the options open to agents: it may include the possibility that each mechanism contributes a pint,

We can now complete our definition of security, by relativising it to a possibility horizon:

SECURITY WITHIN A POSSIBILITY HORIZON: the security of outcome  $O$  in  $w$  at  $t$  within possibility horizon  $H(t)$  is given as follows:

If  $O$  occurs in  $w$ , then  $O$  has *positive* security in  $w$ , and its degree of positive security in  $w$  at  $t$  within  $H(t)$  is given by the distance-at- $t$  between  $w$  and the closest-to- $w$ -at- $t$  world(s) in  $H(t)$  where  $O$  does not occur.

If  $O$  does not occur in  $w$ , then  $O$  has *negative* security in  $w$ , and its degree of negative security in  $w$  at  $t$  in  $H(t)$  is given by the distance-at- $t$  between  $w$  and the closest-to- $w$ -at- $t$  world(s) in  $H(t)$  where  $O$  occurs.

Based on this, we suggest the following completion of our schema:

REASON: you have a teleological reason to  $\varphi$  rather than  $\psi$  at time  $t$ , where  $\varphi$  and  $\psi$  are two mutually incompatible actions or omissions, just in case

- (a) it is an option for you to  $\varphi$  at  $t$ ,
  - (b) it is an option for you to  $\psi$  at  $t$ , and
- there are two incompatible outcomes  $O$  and  $O^*$ , such that
- (c)  $O$  is better than  $O^*$ , and
  - (d)  $O$  is more secure at  $t$  and  $O^*$  is less secure at  $t$  in the closest-to-@-at- $t$  world(s) in  $H(t)$  where you  $\varphi$  at  $t$  than they are in the closest-to-@-at- $t$  world(s) in  $H(t)$  where you  $\psi$  at  $t$ , where  $H(t)$  is the deliberately relevant possibility horizon for the purpose of determining what you have reason to do at  $t$ .

Assuming that (a) it is an option for you to  $\varphi$  at  $t$ , (b) it is an option for you to  $\psi$  at  $t$ , and (c)  $O$  is better than  $O^*$ , REASON entails that the following inferences hold (see Appendix):

THE WHETHER-WHETHER INFERENCE:

If whether  $O$  or  $O^*$  will occur depends on whether you  $\varphi$  or  $\psi$  at  $t$ , then you have a reason to  $\varphi$  rather than  $\psi$  at  $t$ .

THE CAUSAL INFERENCE:

If your  $\varphi$ -ing rather than  $\psi$ -ing at time  $t$  would be a cause of  $O$  rather than  $O^*$ , then you have a reason to  $\varphi$  rather than  $\psi$  at  $t$ .

---

or fails to do so. With this larger possibility horizon, our account delivers the verdict that you have a reason. Nefsky, by contrast, gets the verdict that you have no reason either way.



REASON thus supports the key inferences we identified in Section 3.

## 5. Testing the Account

In this section, we show that REASON delivers intuitively correct results in our three test cases – NUCLEAR SAFETY, THE LAKE, and COORDINATION – as well as in TRAIN TRACKS.

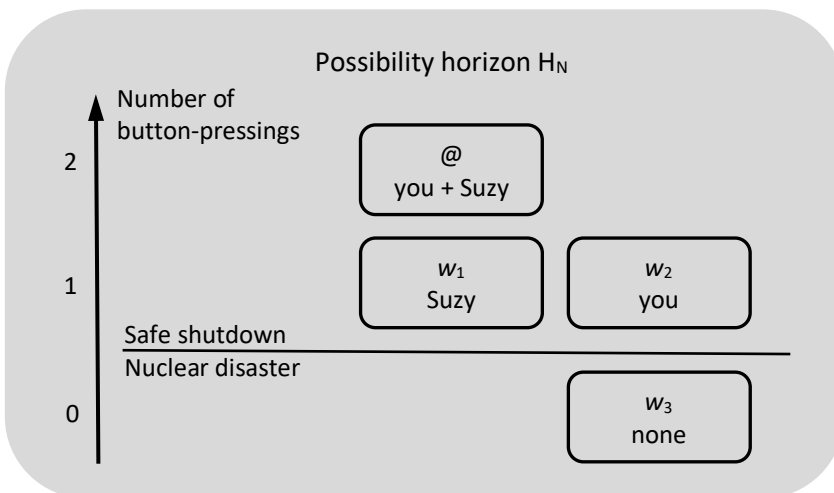
First, consider NUCLEAR SAFETY. Clearly, the first two conditions of REASON are satisfied: (a) it was an option for you to press the safety button at time  $t$ , and (b) it was an option for you not to press the safety button at time  $t$ . Furthermore, (c) is clearly satisfied for the following choice of  $O$  and  $O^*$ :

$O$  = safe shutdown of the reactor

$O^*$  = meltdown of the reactor in a nuclear disaster

From here, we could simply appeal to THE CAUSAL INFERENCE: on the assumption that your pressing your button rather than not is a cause of the reactor's shutting down safely, THE CAUSAL INFERENCE delivers the result that you had a reason to press your button. However, we may also show directly that condition (d) is satisfied:

To do so, we first identify the deliberately relevant possibility horizon. You and Suzy each have two options: pressing your safety button or not. This means that the relevant possibility horizon as a minimum includes  $2^2 = 4$  possible worlds, as illustrated in the following figure:



To see that (d) is satisfied, we need to consider two worlds: the closest-to-@-at- $t$  world within  $H_N$  where you press your button, namely @, and the closest-to-@-at- $t$  world within  $H_N$  where you do not press your button, namely  $w_1$ .

Is the safe shutdown of the reactor *more secure* in @ than in  $w_1$ ? Yes. The safe shutdown of the reactor occurs, and thus has positive security, in both @ and  $w_1$ . However, there is a difference in its degree of security. Relative to @, the closest-at- $t$  world where the reactor is not shut down safely is world  $w_3$ . Relative to  $w_1$ , the closest-at- $t$  world where the reactor is not shut down safely is still  $w_3$ . Clearly, the distance-at- $t$  between @ and  $w_3$  is greater than the distance-at- $t$  between  $w_1$  and  $w_3$ : @ and  $w_1$  both differ from  $w_3$  in terms of whether or not Suzy presses her safety button. But in addition, @ also differs from  $w_3$  in terms of whether or not *you* press your safety button. Thus, the safe shutdown is more secure in @ than in  $w_1$ .

Is the meltdown of the reactor in a nuclear disaster *less secure* in @ than in  $w_1$ ? Yes. The nuclear meltdown does not occur, and therefore has negative security, in both @ and  $w_1$ . Once again, however, there is a difference in degree. The closest-to-@-at- $t$  world where there is a nuclear meltdown is world  $w_3$ , and the closest-to- $w_1$ -at- $t$  world where there is a nuclear meltdown is also world  $w_3$ . As we have seen, the distance-at- $t$  between @ and  $w_3$  is greater than the distance-at- $t$  between  $w_1$  and  $w_3$ . This means that the nuclear meltdown has a higher degree of negative security in @ than in  $w_1$ : the nuclear meltdown is, so to speak, further from happening in @ than in  $w_1$ . Therefore, the nuclear meltdown is *less secure* in @ than it is in  $w_1$ , just as it is *less warm* when the temperature is  $-20^\circ$  than it is when it is  $-10^\circ$ .

This shows that condition (d) is satisfied. Thus, you have a reason to press your button in NUCLEAR SAFETY. A parallel argument shows that Suzy has such a reason too.<sup>17</sup>

Let us next consider THE LAKE. Do you have a reason to use the non-toxic paint rather than the toxic paint at the time  $t$  when you, Vanessa, and Walter are painting your boats?

Clearly, condition (a) and (b) are satisfied: (a) it is an option for you to use the non-toxic paint at  $t$ , and (b) it is an option for you to use the toxic paint. Furthermore, the remaining conditions are satisfied when

O = survival of the ecosystem

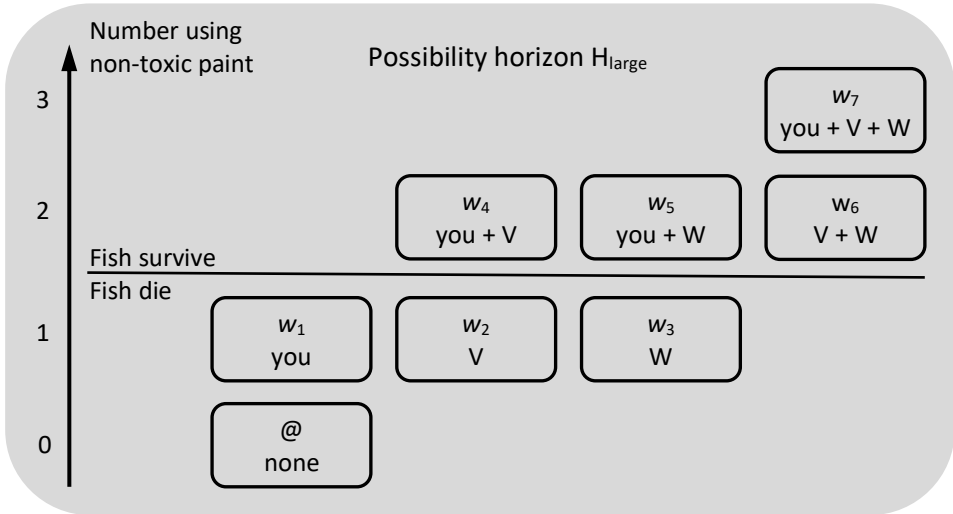
O\* = collapse of the ecosystem

---

<sup>17</sup> If we instead consider the narrow possibility horizon that only includes @ and  $w_1$ , i.e. which holds fixed that Suzy presses her button, REASON delivers the verdict that you did not have a reason to press your button.

Clearly, condition (c) is satisfied: the survival of the ecosystem is better than its collapse.

To show that (d) is satisfied, we first need to identify the deliberatively relevant possibility horizon. There are three agents, each with two options: using the non-toxic paint at  $t$ , or using the toxic paint at  $t$ . Thus, our possibility horizon should as a minimum include every combination of these courses of action, i.e.  $2^3 = 8$  possible worlds, as illustrated below:

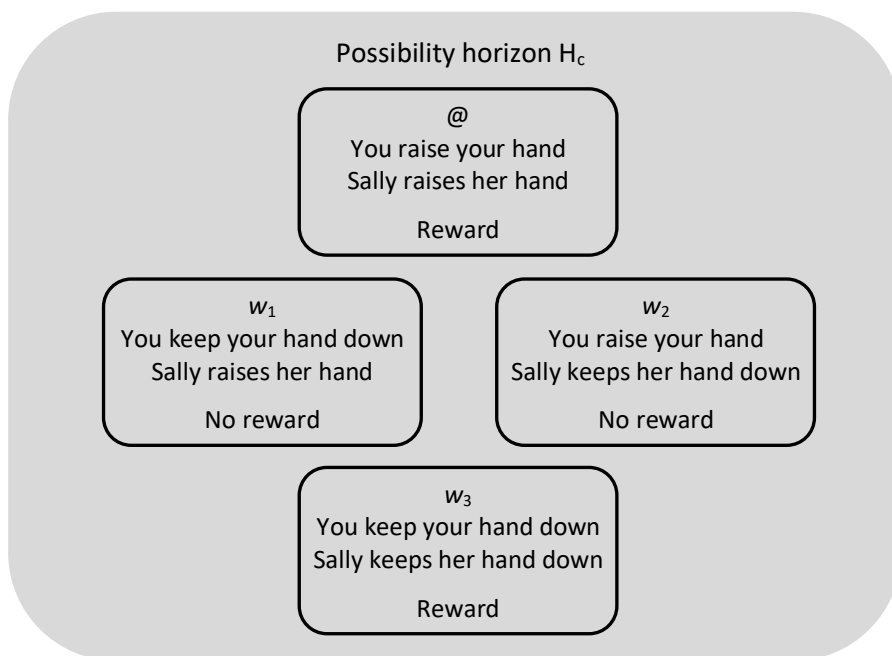


Within this possibility horizon, the closest-to-@-at- $t$  world where you use the non-toxic paint at  $t$  is world  $w_1$ , and the closest-to-@-at- $t$  world where you use the toxic paint is @.

The survival of the ecosystem has negative security in both  $w_1$  and @. The closest-to- $w_1$ -at- $t$  world(s) where the ecosystem survives are  $w_4$ , and  $w_5$ , where you and one other use the non-toxic paint. The closest-to-@-at- $t$  world(s) where the ecosystem survives are  $w_4$ ,  $w_5$ , and  $w_6$ , where two people use the non-toxic paint. The distance-at- $t$  between  $w_1$  and  $w_4$  or  $w_5$  is *smaller* than the distance-at- $t$  between @ and  $w_4$ ,  $w_5$ , or  $w_6$ : to get from  $w_1$  to  $w_4$  or  $w_5$ , only one person needs to change which paint they are using, but to get from @ to  $w_4$ ,  $w_5$ , or  $w_6$ , two people need to change. This means that the survival of the ecosystem is *more secure* in  $w_1$  than it is in @: even though the ecosystem does not survive in either  $w_1$  or @, it is *closer* to surviving in  $w_1$ . A parallel argument shows that the collapse of the ecosystem is *less secure* in  $w_1$ , where you use the non-toxic paint, than it is in @, where you use the toxic one.

Thus condition (d) is satisfied, and we find, as we should, that you had a reason to use the non-toxic paint.

As our third test case, let us consider COORDINATION. Here, we need to verify that REASON delivers the two results we want: first, that you had a reason to raise your hand; and second, that you did *not* have a reason to keep your hand down. To show this, we need to consider the deliberately relevant possibility horizon:



We immediately find that you had a reason to raise your hand rather than keep it down: (a) and (b) are clearly satisfied. Setting

O = each of you getting a million dollars, and

O\* = getting nothing,

condition (c) is also satisfied. By THE WHETHER-WHETHER INFERENCE, we then find that you had a reason to raise your hand: O occurs in the closest-to-@-at- $t$  world where you raise your hand (namely @), while O\* occurs in the closest-to-@-at- $t$  world (namely  $w_1$ ) where you keep it down.

Importantly, REASON also delivers the result that you did *not* have a reason to keep your hand down rather than raising it. For in this case, condition (d) is not satisfied:

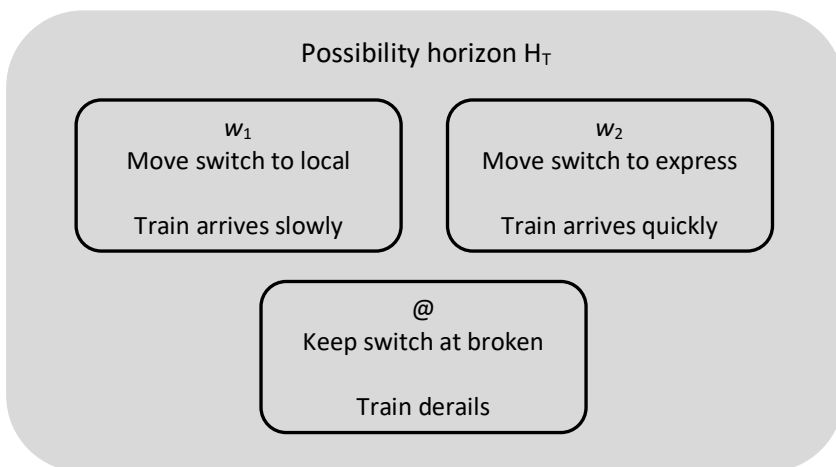
it is not the case that your getting a million dollars each is *more secure* in the closest-to-@-at- $t$  world where you keep your hand down, i.e. in  $w_1$ , than it is in the closest-to-@-at- $t$  world where you raise your hand, i.e. in @. On the contrary, your getting a million dollars each does not occur, i.e. has negative security, in  $w_1$ , while it does occur, i.e. has positive security, in @. Thus, REASON avoids the problem we identified for POTENTIAL CAUSE. It does so by being sensitive to what in fact happened in the actual world – namely, that Sally raised her hand.

In the cases we have considered so far, we found that whenever O was *more secure*, O\* was *less secure*. Indeed, this is so whenever either O or O\* occurs in every world within  $H(t)$  (see the proof of SYMMETRY in Appendix to this chapter). This might make you think that the requirement that O\* should be *less secure* in the closest  $\varphi$ -ing worlds than in the closest  $\psi$ -ing worlds is superfluous. However, this requirement does essential work in delivering the correct result in TRAIN TRACKS:

The challenge in TRAIN TRACKS is to simultaneously accommodate the falsity of (i) and the truth of (ii):

- (i) You have a reason to move the switch to *local* rather than moving it to *express*.
- (ii) You have a reason to move the switch to *local* rather than leaving it at *broken*.

To see how REASON accommodates these verdicts, the first step is to identify the deliberately relevant possibility horizon. This is illustrated below. (We arbitrarily suppose that you leave the switch at *broken*, denoting the world in which you do so “@”. REASON delivers the same results on the supposition that you move the switch to *local* or *express*, so that  $w_1$  or  $w_2$  becomes the actual world. This follows from STABILITY, which we prove in Appendix.)



In the case of both (i) and (ii), we find that there is only one choice of O and O\*, such that condition (c) and the part of (d) that is concerned with O is satisfied:

O = slow arrival

O\* = derailling

Here, (c) is satisfied, since the train's slow arrival is better than its derailling. Furthermore, the train's slow arrival is *more secure* in  $w_1$  than it is in @. The action lies in the part of (d) that is concerned with O\*:

In the case of (i), the relevant comparison is between the closest-to-@-at- $t$  world where you set the switch to *local*, namely  $w_1$ , and the closest-to-@-at- $t$  world where you set the switch to *express*, namely  $w_2$ . And we find that the train's derailling is *just as secure* in  $w_1$  as it is in  $w_2$ . Thus, the part of condition (d) that is concerned with O\* fails to be satisfied, yielding the intuitively correct judgement that (i) is false.

In the case of (ii), there is a different contrast – namely, keeping the switch at *broken*. This means that we have to make a different comparison: the relevant comparison is between  $w_1$  and the closest-to-@-at- $t$  world where you keep the switch at *broken*, namely @. Here, we find that the train's derailling is *less secure* in  $w_1$  than it is in @. Thus, condition (d) is fully satisfied, and we get the intuitively correct judgement that (ii) is true.

This shows that REASON satisfies our desiderata: it delivers intuitively correct results in all our test cases, and, as we have seen above (Section 4), it entails that WHETHER-WHETHER DEPENDENCE and CAUSE give sufficient conditions for having teleological reasons. In the following section we finally show how REASON delivers the intuitively correct result in DROPS OF WATER.

## 6. Drops of Water

In THE LAKE, there is a sharp threshold: if at least two of you use the toxic paint, the ecosystem will collapse; but if at most one uses the toxic paint, the ecosystem will survive. This makes THE LAKE a typical threshold case. In DROPS OF WATER, however, it seems that there is no normatively significant threshold: what matters is how the men in the desert feel, and adding one extra pint to the cart makes no perceptible difference to the suffering of anyone. This feature of DROPS OF WATER has made it very difficult to find a plausible account of reasons for action that can capture the intuition that you do have a reason to donate your pint.

One way to respond to this challenge is to argue that there is a normatively significant threshold after all. For example, Kagan (2011) argues that non-threshold cases are conceptually impossible (for a reply, see Nefsky 2012), and Parfit (1984), Barnett (2018) and Broome (2019) give arguments aiming to show that imperceptible differences do matter morally. REASON offers an alternative response, which does not depend on how this debate turns out:

According to REASON, what matters is that there is a normatively significant difference between outcomes at opposite ends of the spectrum – for example, between the men’s suffering being fully alleviated, and their suffering continuing unmitigated. By donating your pint, you can increase the security of the good outcome that the men’s suffering is fully alleviated, and decrease the security of the bad outcome that the men’s suffering continues unmitigated. Because of that, you have a reason to donate your pint.<sup>18</sup>

In the following, we show in more detail that the conditions of REASON are satisfied. To do so, we consider a particular situation where 6,000 others donate their pints, while you fail to do so. Did you have a reason to donate your pint?

By hypothesis, the first two conditions of REASON are satisfied: (a) it is an option for you to donate your pint, and (b) it is an option for you to keep it to yourself. Further, we may choose O and O\*, such that

O = the full alleviation of the men’s suffering

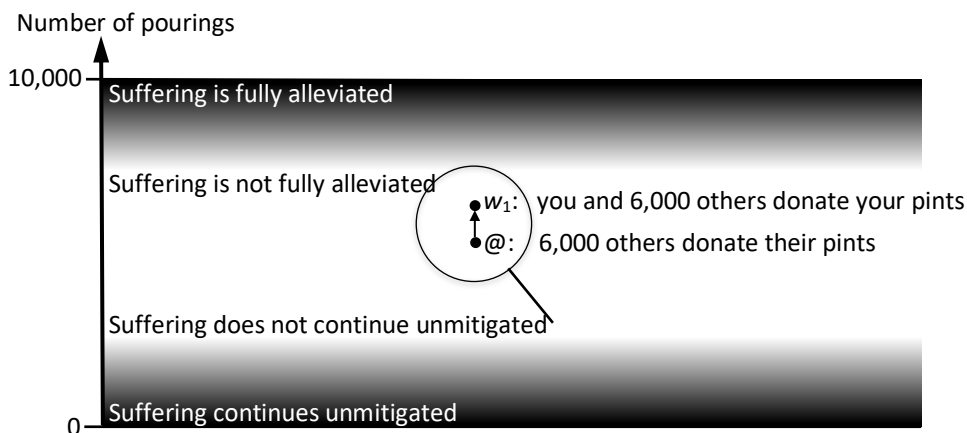
O\* = the unmitigated continuation of the men’s suffering

This ensures that condition (c) is satisfied: the full alleviation of the men’s suffering is clearly better than the unmitigated continuation of their suffering. The key here is that O and O\* are at opposite ends of the spectrum, so that it clearly makes a morally significant difference whether one or the other occurs.

We may now go on to show that (d) is satisfied. To do so, we first need to identify the relevant possibility horizon. You and each of the other 9,999 people can choose between two options: donate your pint, or keep it for yourself. The deliberately relevant possibility horizon therefore has to contain at least  $2^{10,000}$  possible worlds. Of course, we cannot represent all of these worlds individually, but the following illustration will hopefully do:

---

<sup>18</sup> Kagan, Broome, etc., focus on the two adjacent outcomes that would result if you were to either donate your pint or not. By contrast, our approach is more similar to that of Rabinowicz (1989: 39-43), focusing on the whole spectrum of outcomes – including outcomes that would come about if others acted differently.



To ensure that the distance between @ and  $w_1$  is discernible in the figure, we have magnified this part of the figure (as indicated by the stylized magnifying glass). Since there is a vague border between worlds where the men's suffering is fully alleviated, and worlds where it is not, we use a gradient that runs from black to white rather than a sharp line. The same applies to the vague border between worlds where the men's suffering continues unmitigated and worlds where it does not.

Intuitively, it is now clear that (d) the full alleviation of the men's suffering is *more secure* in  $w_1$  than it is in @: if you add your pint, you *take one step closer* to the full alleviation of the men's suffering. And similarly, the unmitigated continuation of the men's suffering is *less secure* in  $w_1$  than it is in @.

However, one might object that it is not obvious that adding a pint to the cart makes the full alleviation of suffering *more secure* (or the unmitigated continuation of suffering *less secure*): one might argue that since it is vague when the men's suffering is fully alleviated, we cannot decide whether the distance between  $w_1$  and the closest-to- $w_1$ -at- $t$  world(s) with full alleviation is in fact shorter than the distance between @ and the closest-to-@-at- $t$  world(s) with full alleviation; just as we cannot decide the exact distance between a point and an interval with fuzzy boundaries. We have two answers to this worry:

First, it seems intuitively clear that adding your pint makes the full alleviation of the men's suffering more secure, and this is so even if it is vague precisely where we reach the full alleviation of suffering: by pouring your pint into the cart, you are taking a step towards the desired outcome. Compare this to being on an airplane that is flying into a cloud: although it is vague precisely when you go from being outside the cloud to being inside the cloud, there may be no doubt about whether you are moving *towards* the cloud.



Second, and more formally, we may look more closely at vagueness. According to an attractive view – presented, for example, by David Lewis – vagueness is semantic indecision:

The only intelligible account of vagueness locates it in our thought and language. The reason it's vague where the outback begins is not that there's this thing, the outback, with imprecise borders; rather there are many things, with different borders, and nobody has been fool enough to try to enforce a choice of one of them as the official referent of the word "outback"

(Lewis 1986c: 213; see also Fine 1975)

We may understand the vagueness of when the men's suffering is fully alleviated in a parallel way: the reason it is vague when the men's suffering is fully alleviated is not that there is this event, the full alleviation of the men's suffering, with imprecise conditions of occurrence; rather there are many events, with different conditions of occurrence, and nobody has been fool enough to try to enforce a choice of one of them as the official referent of "the full alleviation of the men's suffering".

How should we evaluate the truth of a sentence containing vague terms, such as "the full alleviation of the men's suffering is *more secure* in  $w_1$  than it is in  $@$ "? In the following, we will appeal to the supervaluationist proposal (see Fine 1975; Keefe 2000):

**SUPERVALUATIONISM:** A statement is supertrue and therefore true if and only if it is true on all its admissible completely sharp sharpenings.

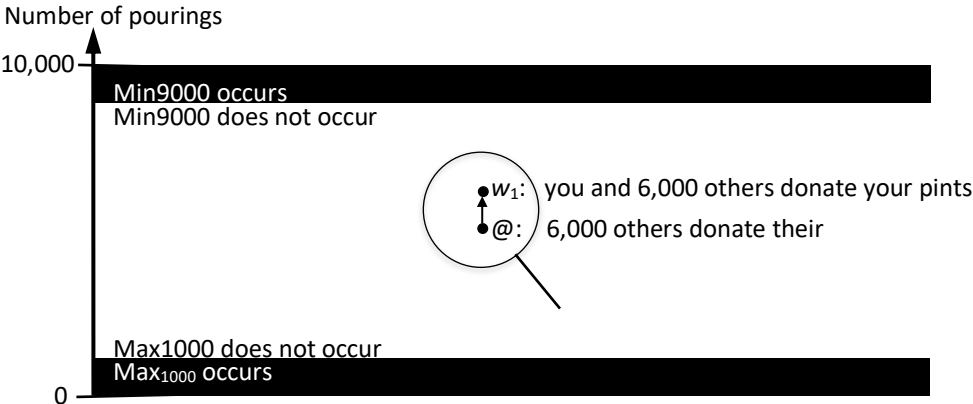
To apply this proposal, the first step is to identify all admissible completely sharp sharpenings. Let  $\text{Exact}(n)$  be the event that occurs just in case each of the ten thousand men is in the brain-state he will be in after drinking if *exactly*  $n$  pints are donated. Let  $\text{Min}(N)$  be the event that occurs just in case  $\text{Exact}(n)$  occurs for some  $n \geq N$ . For an appropriate choice of  $N$ , events such as  $\text{Min}(N)$  look like good candidates for what we mean by "the full alleviation of the men's suffering". For example,  $\text{Min}(9000)$  is an admissible sharpening of "the full alleviation of the men's suffering": it captures both the fact that the men's suffering may be fully alleviated even though a few people fail to donate their pints, and that if the men's suffering is alleviated when 9000 people donate their pints, it is also alleviated when more than 9000 people do so. Similarly, the best candidates for what we mean by "the unmitigated continuation of the men's suffering" are events such as  $\text{Max}(M)$ , where  $\text{Max}(M)$  occurs just in case  $\text{Exact}(n)$  occurs for some  $n \leq M$ . For example,  $\text{Max}(1000)$  is an admissible sharpening of "the unmitigated continuation of the men's suffering".

We may now show that condition (d) is satisfied for any such admissible sharpenings of “the full alleviation of the men’s suffering” and “the unmitigated continuation of the men’s suffering”.<sup>19</sup> For illustration, we consider the following:

$$O = \text{Min}(9000)$$

$$O^* = \text{Max}(1000)$$

It is clear that  $\text{Min}(9000)$  is *more secure* in  $w_1$  than it is in  $@$ . Starting from  $w_1$  (where you donate your pint), only 2,999 people need to act differently in order for  $\text{Min}(9000)$  to occur; but starting from  $@$ , 3,000 people need to act differently. Similarly,  $\text{Max}(1000)$  is *less secure* in  $w_1$  than it is in  $@$ . Starting from  $w_1$ , 5,001 people need to act differently in order for  $\text{Max}(1000)$  to occur; but starting from  $@$ , only 5,000 need to act differently. This is illustrated in the figure below, which indicates the precise conditions of occurrence of  $\text{Min}(9000)$  and  $\text{Max}(1000)$ .



The same argument applies to any admissible sharpening of “the full alleviation of the men’s suffering” and “the unmitigated continuation of the men’s suffering”. We therefore find that all admissible completely sharp sharpenings of

- “the full alleviation of the men’s suffering is *more secure* in  $w_1$  than in  $@$ ”, and
- “the unmitigated continuation of the men’s suffering is *less secure* in  $w_1$  than in  $@$ ”,

<sup>19</sup> Note that condition (c) is also satisfied for any such admissible sharpenings. For example, even when  $\text{Min}(9000)$  occurs in virtue of  $\text{Exact}(9000)$  and  $\text{Max}(1000)$  occurs in virtue of  $\text{Exact}(1000)$ , it is clearly true that the occurrence of  $\text{Min}(9000)$  is better than the occurrence of  $\text{Max}(1000)$ .

are true. From this it follows that these claims are themselves supertrue, and therefore true.<sup>20</sup> REASON thus delivers the intuitively correct result that you have a reason to donate your pint in DROPS OF WATER.

## 7. Conclusion

REASON explains *in virtue of what* you have a teleological reason to donate your pint in DROPS OF WATER: you have such a reason in virtue of the fact that donating your pint makes it *more secure* that the men's suffering will be fully alleviated, and *less secure* that the men's suffering will continue unmitigated.

You might think that REASON comes into conflict with the intuitive thought that only the actual world matters. This thought easily leads to the thought that *mere* differences in security do not matter. However, we think this further step is a mistake: if we disregard mere differences in security, we get an account that is only able to capture how you, *as an individual*, can make a difference to what actually happens. Such an account overlooks the fact that *we together* can make a difference to what actually happens. When this is the case, you have a reason to act when your action *contributes* to our making a difference together.

Putting the point as we just did might evoke top-down accounts of reasons, where your individual reasons derive from what the collective can do. However, REASON captures the intuition that you have a reason to contribute when *we* can make a difference, while being entirely bottom-up. Our rule for determining the deliberatively relevant possibility horizon ensures that we consider all the agents involved in a situation, and every combination of the courses of action that are open to them. When we apply this to DROPS OF WATER, we see that the 10,000 can make a significant positive difference: they can ensure that, in the actual world, the men's suffering is fully alleviated rather than continuing unmitigated. You contribute to making this difference by making the good outcome *more secure*, and the bad outcome *less secure* – that is, by donating your pint rather than keeping it to yourself.

---

<sup>20</sup> We would get the same result if we instead appealed to an epistemic view of vagueness according to which vagueness is a kind of ignorance. On such a view, “the full alleviation of the men's suffering” refers to one event with precise conditions of occurrence; we just don't know which (see Williamson 1994). The same applies to “the unmitigated continuation of the men's suffering”. Presumably, these events have the form  $\text{Min}(N)$  and  $\text{Max}(M)$ . From the reasoning above, it now immediately follows that (c) and (d) are true on an epistemic view.

# Appendix

In this Appendix, we prove four results. We begin with an auxiliary result that we call STABILITY, namely that whether you have a reason to  $\varphi$  rather than  $\psi$  at  $t$  does not depend on what you actually do at  $t$ . Next, we prove that THE WHETHER-WHETHER INFERENCE and THE CAUSAL INFERENCE hold. Finally, we prove SYMMETRY, namely: whenever either O or O\* occurs in every world within  $H(t)$ , it is the case that if O is *more secure* in the closest  $\varphi$ -ing world than in the closest  $\psi$ -ing world, then O\* is *less secure* in the closest  $\varphi$ -ing world than in the closest  $\psi$ -ing world.

## Stability

REASON entails what we call STABILITY:

STABILITY: whether or not you have a reason to  $\varphi$  rather than  $\psi$  at  $t$  does not depend on what you actually do at  $t$ .

This is an intuitively pleasing result: it fits the thought that, although you may need to know lots of other things about the actual world in order to figure out whether you have objective reason to  $\varphi$  rather than  $\psi$  at  $t$ , you do not need to know what you will in fact do at  $t$ .

Consider REASON. The only condition that references @ is (d), which requires us to look at the closest-to-@-at- $t$  world where you  $\varphi$  at  $t$ , and the closest-to-@-at- $t$  world where you  $\psi$  at  $t$ . These worlds are the same no matter what you do. To see this, suppose that you have the following range of options:  $\varphi$ ,  $\psi$ , or  $\chi$  (including further options does not change the structure of the argument). Since  $\varphi$ ,  $\psi$ , or  $\chi$  are all of your options, you do one of these in the actual world. It now follows from our construction of the deliberately relevant possibility horizon  $H(t)$  that  $H(t)$  includes three worlds –  $w_\varphi$ ,  $w_\psi$ , and  $w_\chi$  – that are exactly alike, except that you do  $\varphi$  in  $w_\varphi$ ,  $\psi$  in  $w_\psi$ , and  $\chi$  in  $w_\chi$ , and where one of these three worlds is @. Irrespective of what you do – that is, irrespective of which of the three worlds that is @ – we find that the closest-to-@-at- $t$  world where you  $\varphi$  is  $w_\varphi$ , and the closest-to-@-at- $t$  world where you  $\psi$  is  $w_\psi$ . It therefore makes no difference to the verdict of REASON whether you do  $\varphi$ ,  $\psi$ , or  $\chi$ .

## The Whether-Whether Inference

REASON entails that, when (a) it is an option for you to  $\varphi$  at  $t$ , (b) it is an option for you to  $\psi$  at  $t$ , and (c) O is better than O\*, the following holds:

THE WHETHER-WHETHER INFERENCE:

If *whether* O or O\* will occur depends on *whether* you  $\varphi$  or  $\psi$  at  $t$ , then you have a reason to  $\varphi$  rather than  $\psi$  at  $t$ .

More carefully, the antecedent is satisfied when the following holds: in the closest world(s) in  $H(t)$  where you  $\varphi$  at  $t$ , O occurs; and in the closest world(s) in  $H(t)$  where you  $\psi$  at  $t$ , O\* occurs.

Suppose that this is the case. If so, O has positive security in the closest  $\varphi$ -ing world(s), and negative security in the closest  $\psi$ -ing world(s). It immediately follows that (d) O is *more secure* and O\* is *less secure* in the closest  $\varphi$ -ing world(s) than they are in the closest  $\psi$ -ing world(s). Since condition (a), (b), and (c) are also satisfied, REASON entails that you have a reason to  $\varphi$  rather than  $\psi$  at  $t$ .

## The Causal Inference

REASON also entails that, when (a) it is an option for you to  $\varphi$  at  $t$ , (b) it is an option for you to  $\psi$  at  $t$ , and (c) O is better than O\*, the following holds:

THE CAUSAL INFERENCE:

If your  $\varphi$ -ing rather than  $\psi$ -ing at time  $t$  *would* be a cause of O rather than O\*, then you have a reason to  $\varphi$  rather than  $\psi$  at  $t$ .

More carefully, the antecedent is satisfied when the following holds: your  $\varphi$ -ing rather than  $\psi$ -ing at time  $t$  *would* be a cause of O rather than O\* within  $H(t)$ .

Suppose first that the antecedent is satisfied, and that you  $\varphi$  in @. Then your  $\varphi$ -ing rather than  $\psi$ -ing *is* a cause of O rather than O\* in  $H(t)$ . As we claimed in Section 4, (d) is a necessary condition for causation. Thus, condition (d) is satisfied.

Suppose next that the antecedent is satisfied, and that you do not  $\varphi$  in @. Let  $w_\varphi$  be the closest-to-@-at- $t$  world in  $H(t)$  where you  $\varphi$ . From the assumption that your  $\varphi$ -ing rather than  $\psi$ -ing *would* be a cause of O rather than O\* in  $H(t)$ , it follows that in  $w_\varphi$  your  $\varphi$ -ing rather than  $\psi$ -ing *is* a cause of O rather than O\* in  $H(t)$ . Thus, condition (d) is satisfied in  $w_\varphi$ . From STABILITY, it now follows that it is satisfied in @ as well.

## Symmetry

Finally, we show that REASON entails what we call SYMMETRY:

SYMMETRY: Suppose that either O or O\* occurs in every world in H(t). Then the following holds for any two world worlds  $w_\phi$  and  $w_\psi$  within H(t): if O is *more secure* in  $w_\phi$  than it is in  $w_\psi$ , then O\* is *less secure* in  $w_\phi$  than it is in  $w_\psi$ .

Consider an arbitrary world  $w$  within H(t). Since either O or O\* occurs in every world in H(t), either O or O\* occurs in  $w$ , while the other does not (since they are incompatible). The event that occurs in  $w$  (e.g. O) has positive security in  $w$ , and its degree of positive security is given by distance-at- $t$  between  $w$  and the closest-to- $w$ -at- $t$  world(s) in H(t) where the other event (e.g. O\*) occurs instead. The event that does not occur in  $w$  (e.g. O\*) has negative security in  $w$ , and its degree of negative security is given by the distance-at- $t$  between  $w$  and the closest-to- $w$ -at- $t$  world(s) in H(t) where it (e.g. O\*) does occur. Now, if either O or O\* occurs in every world in H(t), the closest-to- $w$ -at- $t$  world(s) in H(t) where one event (e.g. O) does not occur *just are* the closest-to- $w$ -at- $t$  worlds where the other event (e.g. O\*) occurs. Thus, the distance that determines the occurring event's degree of positive security in  $w$  also determines the non-occurring event's degree of negative security in  $w$ . From this, it follows that for any two worlds  $w_\phi$  and  $w_\psi$ , it holds that if O is *more secure* in  $w_\phi$  than it is in  $w_\psi$ , then O\* is *less secure* in  $w_\phi$  than it is in  $w_\psi$ .



## 6. Using REASON

In Chapters 2, 3 and 4, I considered a number of solutions to the inefficacy argument, and argued that they generate counterexamples. In Chapter 5, Caroline Touborg and I showed that REASON gives the intuitively correct verdict in some of these cases. It gives the right verdict in collective impact cases with a threshold (sometimes called overdetermination cases), like NUCLEAR SAFETY and THE LAKE, in collective impact cases without a threshold, like DROPS OF WATER, in coordination games like COORDINATION, and in cases where it matters what the relevant contrast is, like TRAIN TRACKS. Still, you might wonder whether REASON delivers the intuitively correct verdict in the other cases I have discussed. In this chapter, I will go through these cases. I will argue that REASON gives the right verdict in switching cases, in early and late pre-emption cases, in the case of climate change, in double prevention cases, in cases of transitivity failure, in Julia Nefsky's (2021) VENDING MACHINE case, and in the variant of DROPS OF WATER where the cart is already full when you arrive with your pint.

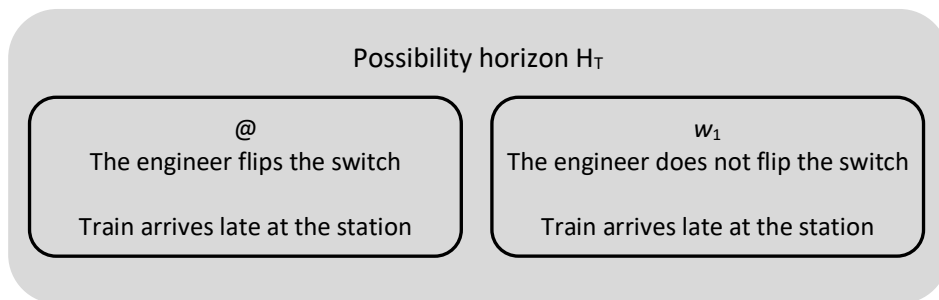
### Switching Cases

Richard Wright's (1985, 2013) NESS account of causation leads to counterintuitive results in switching cases. I considered THE ENGINEER (p. 27 and p. 90), where an engineer flips a switch so that a train travels down the right-hand track instead of the left-hand one, but where the tracks reconverge so that the train arrives at its destination at the time it would have arrived anyway (a little late) had it not been diverted. Here, it seems that the engineer's flipping the switch was not a cause of the train's arriving late. However, NESS entails that it is. The engineer's flipping of the switch was necessary for the sufficiency of an existing antecedent set that was sufficient for the train's arrival at its destination. This point is about causation, not about outcome-related reasons. Still, since NESS gives counterintuitive verdicts on causation in some cases, any account of reasons that builds on this account is likely to sometimes give counterintuitive verdicts about reasons. Our account does not build on NESS, and can deliver the intuitively correct verdict on the engineer's reasons.

In THE ENGINEER, it seems that the engineer has no outcome-related reason to flip the switch. REASON captures this. To see this clearly, we first have to clarify the



relevant possibility horizon. In this case, there are two possibilities: either the engineer flips the switch or she does not. If she flips the switch ( $\varphi$ ), the train will arrive a little late at the station (outcome O). If she does not flip the switch ( $\psi$ ), the train will arrive at the station at the same time (O\*). So, we get the following possibility horizon:



In this case, condition (c) of REASON is not satisfied. Neither of outcomes O and O\* is better than the other. Therefore, REASON does not entail that the engineer has an outcome-related reason to flip the switch.

One might also wonder whether condition (d) of REASON is satisfied. Above, I treated O and O\* as two different but indistinguishable outcomes. If you think that O and O\* are different outcomes, (d) is satisfied. O is more secure and O\* less secure in the closest-to- $@$ -at- $t$  world where the engineer flips the switch (i.e.  $@$ ) than they are in the closest-to- $@$ -at- $t$  world where she does not (i.e.  $w_1$ ). However, there is really just one outcome in this case: the train's late arrival at the station. O and O\* denote one and the same outcome. If you look at the case this way, (d) is not satisfied. O is not more secure in the closest-to- $@$ -at- $t$  world where the engineer flips the switch ( $@$ ) than it is in the closest-to- $@$ -at- $t$  world where she does not ( $w_1$ ). Rather, O is equally secure in both worlds. Anyway, the result is that at least one of REASON's conditions is not satisfied, which means that REASON correctly entails that the engineer lacked an outcome-related reason to flip the switch.

To be precise, REASON entails that the engineer lacked an outcome-related reason to flip the switch rather than to leave it in place. In this chapter, I will frequently omit to mention the relevant contrast for the sake of simplicity. I hope that it will be clear enough what the relevant contrast is anyway.

## Early and Late Pre-emption Cases

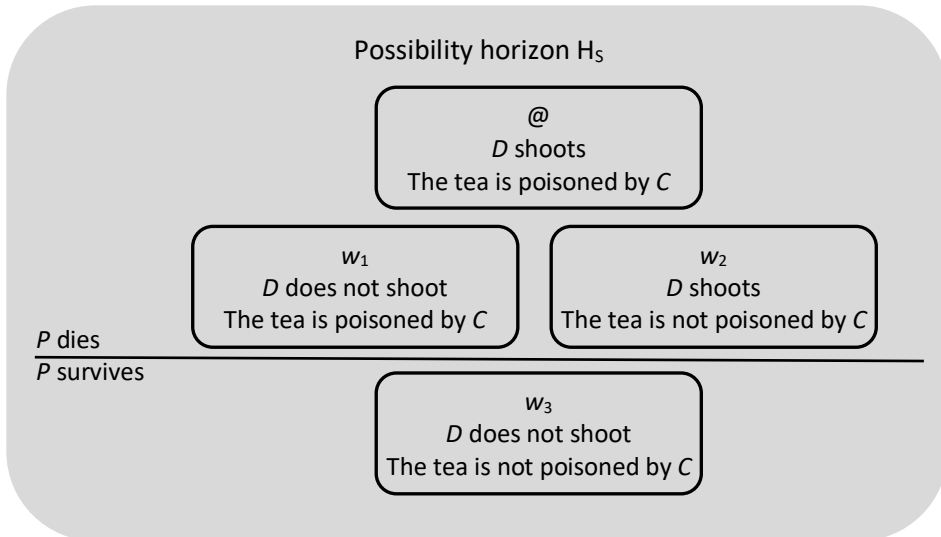
In Chapter 3, we saw that David Lewis' (1973a, 1986a, 1986b) early analysis of causation gave counterintuitive verdicts in late pre-emption cases like SHOOTING AND POISONING.

[SHOOTING AND POISONING:] *D* shoots and kills *P* just as *P* was about to drink a cup of tea that was poisoned by *C*.

(Wright 1985: 1775)

In this case, it seems that *D*'s shooting was a cause of *P*'s death. However, Lewis' early account of causation entails that it was not. Therefore, any account of outcome-related reasons that builds on this account is likely to give mistaken verdicts about reasons.

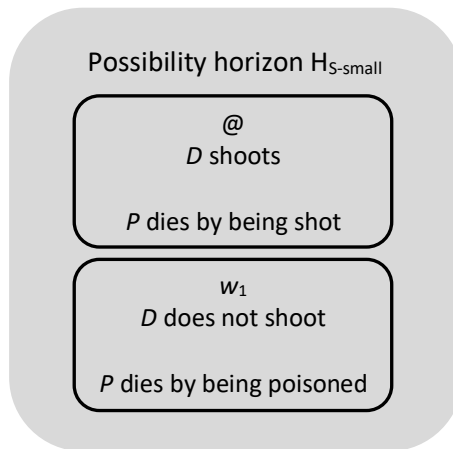
Intuitively, *D* had a survival-of-*P*-related reason not to shoot *P*, at least if we assume that *P*'s death is a bad thing. This is also what REASON entails. As usual, to see this, we have to begin by identifying the relevant possibility horizon. There are four possibilities at the time *t* at which *D* shoots *P*, as shown in the following possibility horizon:



Now we can see that all four conditions of REASON are satisfied. First, there are worlds within  $H_s$  where *D* shoots, and worlds where *D* does not. That is, it is (a) an option for *D* to shoot at *t* and (b) an option for *D* not to shoot at *t*. Second, it is better

that  $P$  survives than that  $P$  dies (as we assume), so (c) is also satisfied. Finally,  $P$ 's survival is more secure and  $P$ 's death is less secure in the closest-to- $@$ -at- $t$  world where  $D$  does not shoot ( $w_1$ ) than they are in the closest-to- $@$ -at- $t$  world where  $D$  does shoot ( $@$ ). In  $w_1$  the only thing that needs to change in order for  $P$  to survive is  $C$ 's poisoning the tea, while in the actual world  $@$   $P$  will only survive if  $C$  does not poison the tea and  $D$  does not shoot  $P$ . In other words,  $P$ 's survival is further from happening in  $@$  than it is in  $w_1$ . Therefore, (d) is also satisfied, and REASON correctly entails that  $D$  has a survival-of- $P$ -related reason not to shoot.

Admittedly, the intuition that  $D$  has an outcome-related reason not to shoot  $P$  is not entirely obvious. You may think, for instance, that  $D$  lacks that reason, since it is already guaranteed that  $P$  will die. REASON can explain this intuition as well. The idea that it is guaranteed that  $P$  will die amounts to the claim that there is no possibility that  $P$  will survive. This gives us a different, smaller, possibility horizon. Given that  $C$  has poisoned the tea and that  $P$  is going to drink it, there are only two possibilities,  $P$  dies by being shot and  $P$  dies by being poisoned.



Here, as long as we assume that it does not matter whether  $P$  dies by being shot or by being poisoned, REASON will deliver the result that  $D$  lacks an outcome-related reason to refrain from shooting  $P$ . He might have other reasons not to shoot  $P$  – for instance that he does not want to be the one who pulls the trigger – but he does not have an outcome-related reason.

Further, if we instead think that it is better for  $P$  to die by being shot than it is for  $P$  to die by being poisoned – for instance because it is better for  $P$  to die quickly and almost painlessly than it is for  $P$  to die a slow, agonising death – REASON entails that  $P$  has an outcome-related reason to shoot  $P$ . There is an option for  $D$  to shoot, and an option not to shoot, so conditions (a) and (b) are satisfied. Further, that  $P$

dies by being shot is better than that  $P$  dies by being poisoned, so (c) is also satisfied. Finally, that  $P$  dies by being shot is more secure and that  $P$  dies by being poisoned is less secure in the closest-to-@-at- $t$  world where  $D$  shoots (@) than they are in the closest-to-@-at- $t$  world where  $D$  does not shoot ( $w_1$ ): in @  $P$  dies by being shot and in  $w_1$   $P$  dies by being poisoned. Hence, condition (d) is also satisfied, meaning that  $D$  has an outcome-related reason to shoot  $P$ .

As will have become clear by now, REASON gives different verdicts depending on which possibilities we take to be relevant. In one respect, this is not a problem. REASON gives the intuitively correct verdict given a certain view of the case. This is exactly what we wanted: we wanted a principle that can explain our intuitions about any given case, even if those intuitions change depending on which possibilities we take to be relevant. In another respect, this is a problem. We might want a definitive answer about what outcome-related reasons there are in any given case. In chapter 11, it is argued that in some cases, there are considerations that licence us to think that a certain possibility horizon is the correct one. For now, I will set this issue aside.

A final observation I wish to make here is that REASON treats early and late pre-emption cases in the same way. Therefore, and for the sake of brevity, I will skip discussing the early pre-emption case we have considered (that is WINDOW BREAKING: see p. 76).

## Climate Change

Does REASON give the intuitively right verdict about the climate-change-related reasons we have? I think it does. Climate change has been described as an overdetermination case (Cripps 2013), a pre-emption case (Lawford-Smith 2016; Eriksson 2019), a collective impact case with a threshold (Kagan 2011),<sup>1</sup> a collective impact case without a threshold (Nefsky 2012; Kingston & Sinnott-Armstrong 2018; Nefsky 2019),<sup>2</sup> and a case where each act does make a difference to the outcome (Broome 2019). The short explanation of why REASON gives the intuitively right verdict about the reasons we have to mitigate climate change is that it gives intuitively right verdicts in all these kinds of case. While Shelly Kagan (2011) has to prove that climate change is a threshold case in order to be able to explain the reasons intuition in this case, REASON gives the right verdict regardless of whether climate change is categorised as a threshold case (more about this in Chapter 8). And, while Anton Eriksson has to show that climate change is a case of early pre-

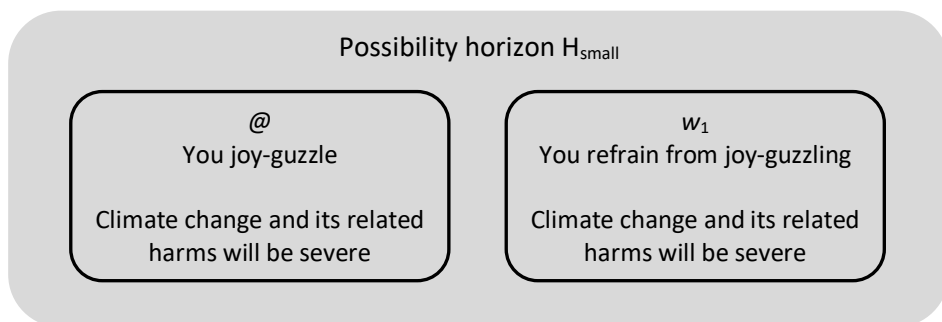
---

<sup>1</sup> Collective impact cases with a threshold and overdetermination cases are the same cases.

<sup>2</sup> To be more precise, Nefsky does not say that climate change is a non-threshold case. Rather, she says that we cannot exclude that it is.

emption rather than late pre-emption or overdetermination, REASON gives the right verdict regardless of which of these options is taken. Further, while John Broome has to prove either that climate change is a case where each act makes a difference to which climate-change-related harms occur or that imperceptible harms are morally relevant, REASON gives the intuitively correct verdict whether or not a single drive makes a difference for the climate, and whether or not imperceptible harms are harms (more about this in Chapters 7 and 9).

In addition, REASON can explain our torn intuitions about what reasons we have in relation to climate change. We might, for instance, rationalise matters along the following lines: given that others will continue using their fossil fuel powered cars, climate change and its related harms will be just as severe whether I go joy-guzzling or not, so I might as well go joy-guzzling. When we rationalise in this way, we tacitly affirm the antecedent (“Given that others will use their fossil fuel cars...”), which amounts to treating what others do as fixed. This confines us to a small possibility horizon only containing two possibilities at the time  $t$  when the option of joy-guzzling is being considered:

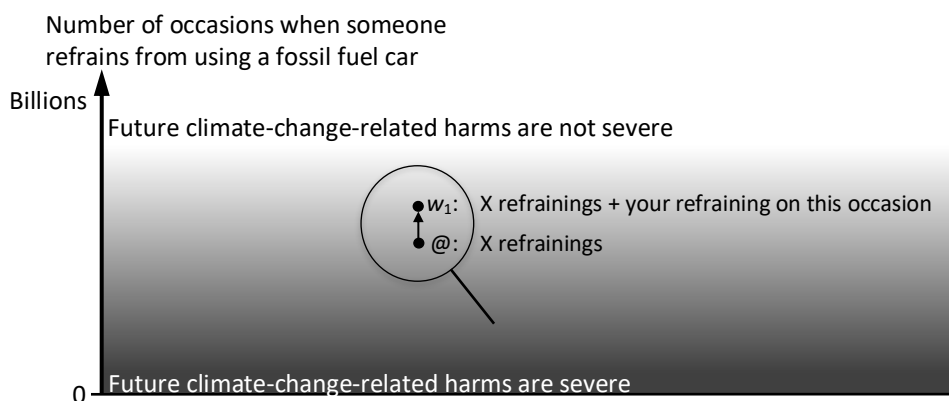


Here, I have arbitrarily set the actual world as the one where you joy-guzzle.<sup>3</sup> Given this limited possibility horizon, REASON entails that you lack a climate-change-related reason to refrain from joy-guzzling. By hypothesis, the outcome will be just the same whether you joy-guzzle or not ( $O$  and  $O^*$  are the same outcomes), so condition (c) of REASON is not satisfied. Moreover, this outcome is just as secure whether or not you joy-guzzle, which means that (d) is not satisfied either. The reasoning here is similar to that concerning SHOOTING AND POISONING and the smaller possibility horizon  $H_{S\text{-small}}$ .

<sup>3</sup> It does not matter which world you set as the actual one. See STABILITY, which you find in the appendix of the previous chapter.

However, we might think about climate change in a different way. We might assume that it is possible to avoid some future climate-change-related harms if enough people refrain from using fossil fuel cars, and that as a result there is a climate-change-related reason not to joy-guzzle. Rationalising matters along these lines, we do not treat what others do as fixed. Each of the others here has a choice of either using fossil fuel cars or refraining from doing so. If enough of them use fossil fuel cars on enough occasions, future climate change and its related harms will be more severe than they would have been if enough of them had refrained from doing so on enough occasions. Rationalising along these lines, we work with a less limited possibility horizon, containing various combinations of possible choices.

### Possibility horizon $H_{large}$



I have arbitrarily set the actual world @ as a world where you do not refrain from joy-guzzling. X is the sum of all the times, in the actual world, that someone refrains from using a fossil fuel powered car when presented with the option of going for a ride.

Given this larger possibility horizon, REASON entails that you have an outcome-related reason to refrain from joy-guzzling. You have the option of refraining from joy-guzzling, and you have the option of going joy-guzzling. So, condition (a) and (b) are satisfied. Further, (c) it is better if future climate-change-related harms are not severe than it is if they are severe. Finally, the less than severe future climate-change-related harms are more secure and the severe future climate-change-related harms are less secure in the closest-to-@-at-t world where you refrain from joy-guzzling (i.e.  $w_1$ ) than they are in the closest-to-@-at-t world where you joy-guzzle

(i.e. @). In  $w_1$ , one person fewer is required to refrain from joy-guzzling (or from using a fossil fuel car in some other way) in order for the future climate-change-related harms not to be severe. The reasoning here is similar to that concerning DROPS OF WATER in the previous chapter.

This means that REASON can explain both the intuition that you lack a climate-change-related reason to refrain from joy-guzzling and the intuition that you have such a reason. The fundamental determinant is how you think of the case – or, in more theoretical language, what possibility horizon you adopt. If you treat what others do as fixed, you lose sight of the possibility that severe climate change might be avoided, with the result that you do not see any reason to reduce your own emissions. But if you treat it as an open possibility that others reduce their emissions, the possibility that severe climate change can be avoided comes into view, as does a reason to reduce your emissions. The question then becomes: How should you view the situation? Should you treat what others do as a fixed background condition, or should you treat it as an open possibility that they act otherwise? Again, I will set this issue aside for now. I return to it in Chapter 11, where, together with Touborg, I argue that the larger possibility horizon is the more accurate one.

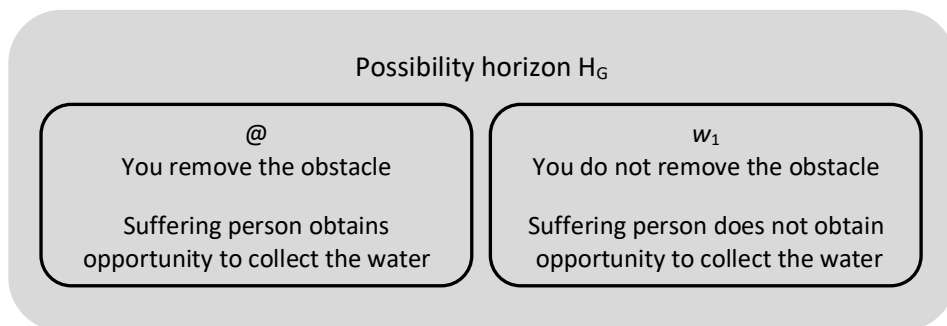
## Double Prevention Cases

I will now turn to a less complex case. You might think that a cause must be connected to its effect via a physical process – that atoms and molecules must bump into each other and exert forces on each other. However, this assumption about causation leads to counterintuitive verdicts in double prevention cases. In Chapter 4, I considered the following case:

GUTTER: There is one person in a nearby desert suffering from thirst. You are standing in front of a long gutter leading to this person. There is no one else around. You know that one pint of water will come flowing through the gutter soon. There is a removable obstacle in the gutter right in front of you. If the obstacle is removed, the person suffering from thirst will be able to collect the pint of water at the end of the gutter. If the obstacle is left in place, the water will not reach the person suffering from thirst. Instead, it will overflow, allowing you to collect it in an empty glass of yours.

If you remove the obstacle, it seems that you are causing the suffering person's opportunity to collect the water. However, if we hold on to the idea that a cause must be connected to its effect via a physical process, it follows that you are not causing this. Therefore, any account of outcome-related reasons that builds on this idea is likely to give counterintuitive verdicts about reasons.

REASON, however, gives the right verdict that you have an outcome-related reason to remove the obstacle. As usual, we have to first decide the relevant possibility horizon at the time  $t$  when you have the option of removing the obstacle. Here, I have arbitrarily set the actual world as the world where you do remove the obstacle.



In  $H_G$ , whether the suffering person obtains an opportunity to collect the water counterfactually depends on whether you remove the obstacle, and it is better that the suffering person obtains the opportunity than it is that he does not. So, REASON quite straightforwardly entails that you have a reason to remove it. Here, I am leaning on THE WHETHER-WHETHER INFERENCE, shown in the appendix to the previous chapter, which says that if some better outcome will occur if you  $\phi$ , and some worse outcome will occur if you do not  $\phi$ , you have a reason to  $\phi$  (and this is entailed by REASON).

## Cases of Transitivity Failure

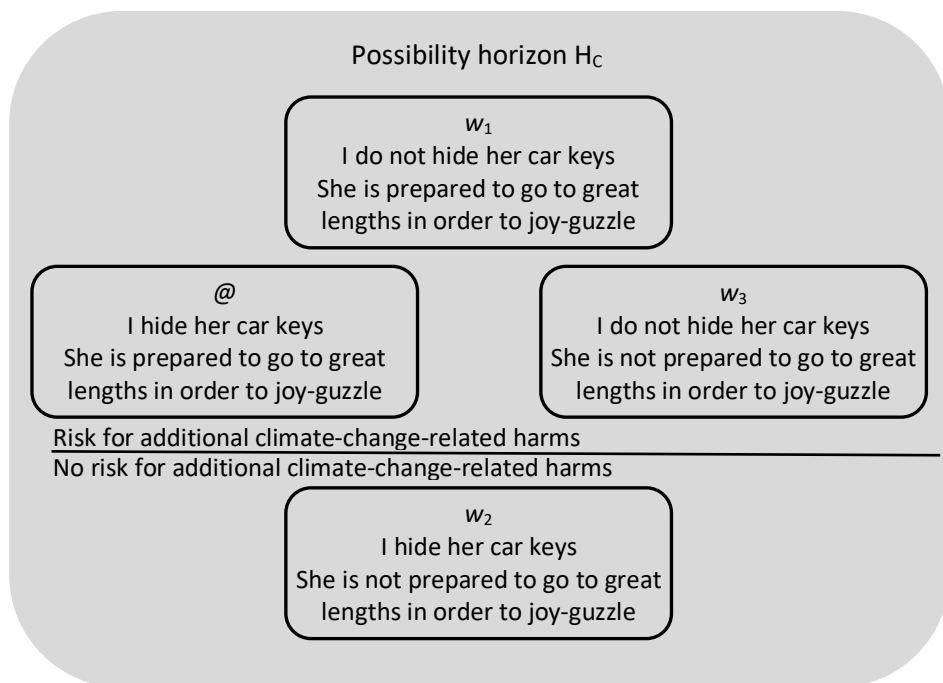
Eriksson's (2019) account of why you have reasons to reduce your emissions of greenhouse gases, which builds on Lewis' (1973a, 1986a, 1986b) early account of causation, gives counterintuitive verdicts in cases of transitivity failure. We have already considered the following case:

**CAR KEYS:** One Sunday morning, I hide my friend's car keys in the hope of making her come along for a bike ride instead of going joy-guzzling as she usually does. However, she manages to hot-wire her car, and goes joy-guzzling anyway.



Lewis' account entails that I caused my friend's leisure drive on this occasion. Eriksson builds on Lewis' account of causation, and claims that I have reasons not to cause harm. We can assume with Eriksson that a single leisure drive with a gas-guzzling car has some probability of triggering climate-change-related harms. Although this assumption requires careful defence, this is not the issue here. The problem is that if, like Eriksson, we apply Lewis' account of causation, we have to conclude that I have climate-change-related reasons not to try to make my friend tag along for a bike ride by hiding her car keys. This, surely, is the wrong verdict.

REASON, however, gives the right verdict about my outcome-related reasons in this case. There are four relevant possibilities at the time  $t$  at which I have the option of hiding my friend's car keys, as indicated in the following possibility horizon:



In all these worlds, I assume, my friend intends to go joy-guzzling at the time when I might hide her car keys. The question is how determined she is to do so, and whether I try to hinder her from going for this drive by hiding her car keys.

REASON entails that I have a climate-change-related reason to hide the keys. First, (a) I have an option of either hiding them or (b) refraining from doing so. Next, (c)

it is better that there is no risk of additional climate-change-related harms than that there is such a risk. Finally, there being no risk of additional climate-change-related harms is more secure and there being such a risk is less secure in the closest-to-@-at- $t$  world where I hide her car keys (@) than they are in the closest-to-@-at- $t$  world where I do not ( $w_1$ ). In @, only one thing needs to change in order for there to be no risk of additional climate-change-related harms: her preparedness to go to great lengths in order to joy-guzzle. However, in  $w_1$ , two things need to change in order for there to be no such risks: I must hide her car keys, and she must not be prepared to go to great lengths in order to joy-guzzle. This means that (d) is satisfied, and thus we can conclude that REASON delivers the intuitively correct verdict that I have a climate-change-related reason to hide my friend's car keys in this case.<sup>4</sup>

## Superfluous Contributions to the Underlying Dimension

Cases like the following pose a problem for Wieland and van Oeveren's (2020) account of outcome-related reasons:

VENDING MACHINE: A, B, and C are walking in a national park, when they come across two hikers who have been lost for days in the backcountry. They are starving. Luckily, there is a vending machine nearby, selling granola bars for \$4 each. The machine accepts all coins and bills, but it does not give change. The two starving hikers do not have any money. But A has a \$5 bill, B has a \$10 bill, and C has a quarter. There is no one else around.

(Nefsky 2021)<sup>5</sup>

Here, intuitively, C has no food-for-the-hikers-related reason to put his quarter into the vending machine. There is no way in which his doing so would contribute to the hikers' getting something to eat. A and B, on the other hand, have food-for-the-hikers-related reasons to put their money into the vending machine.

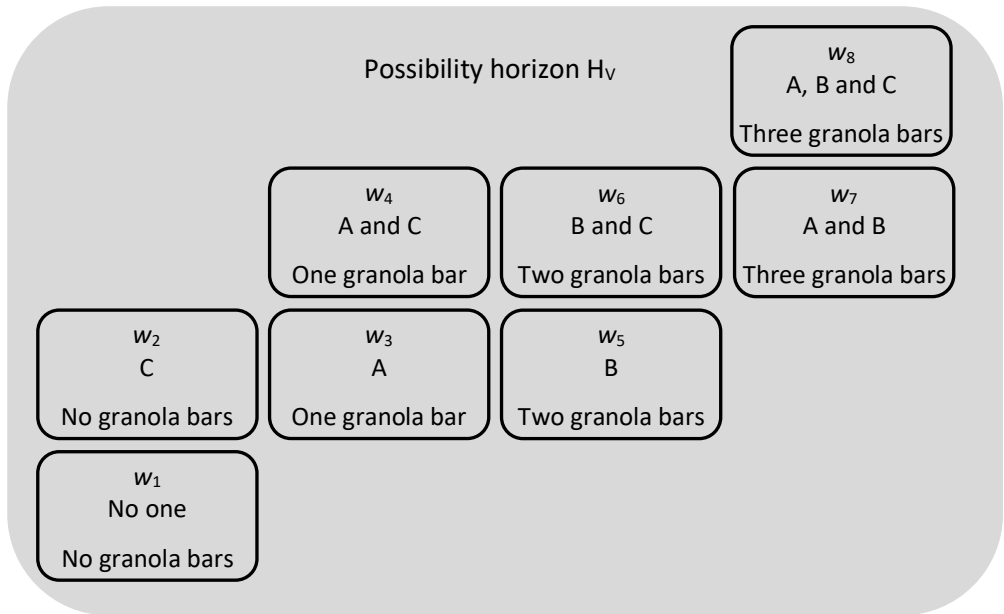
Does this example also pose a problem for REASON? Does REASON entail that C has a reason to put his quarter into the vending machine because his doing so increases the security of the desired outcome that the two hikers get something to eat?<sup>6</sup> The answer is that REASON does not entail this. To see this, we first have to settle on the relevant possibility horizon, at time  $t$ , when C could put his quarter into the vending machine. There are eight possibilities in this case, as follows:

---

<sup>4</sup> For a parallel case, see Touborg (2018: 229-33).

<sup>5</sup> Nefsky (2015) considers a similar example, featuring car-pushing.

<sup>6</sup> Gunnar Björnsson has raised this worry in discussions.



Let us suppose that the actual world  $@$  is  $w_1$  – a world in which no one puts money into the machine. We can now ask whether REASON entails that C has an outcome-related reason to put his quarter into the vending machine. It does not. In the closest-to- $@$ -at- $t$  world where C puts his money into the vending machine ( $w_2$ ), the outcome that the hikers will get no granola bar is just as secure as it is in  $@$ . To see this, we have to compare the distance between  $@$  and the closest-to- $@$ -at- $t$  world(s) where the hikers get something to eat with the distance between  $w_2$  and the closest-to- $w_2$ -at- $t$  world(s) where the hikers get something to eat. At this point, it matters whether it is A or B who is the more reluctant to donate money. If A is less reluctant than B, the closest-to- $@$ -at- $t$  world(s) where the hikers get something to eat is  $w_3$ , and the closest-to- $w_2$ -at- $t$  world(s) where the hikers get something to eat is  $w_4$ . Therefore, the distance between  $@$  and the closest-to- $@$ -at- $t$  world(s) where the hikers get something to eat is the distance between  $@$  and  $w_3$ , and the distance between  $w_2$  and the closest-to- $w_2$ -at- $t$  world(s) where the hikers get something to eat is the distance between  $w_2$  and  $w_4$ . These two distances are the same. What separates  $@$  from  $w_3$  is that A donates his money in one world but not in the other, and what separates  $w_2$  and  $w_4$  is also that A donates his money in one world but not in the other. Therefore, the outcome that the hikers get no granola bars is exactly as secure in  $@$  as in  $w_2$ . This means that condition (d) of reason is not satisfied, and this entails in turn that C lacks an outcome-related reason to put his quarter into the slot in the vending machine. Parallel arguments can be made in the case where B is less reluctant than A to donate his money, and in the case where A and B are equally reluctant to donate

their money. So far, we have only considered what REASON entails if we suppose that the actual world is  $w_1$ . However, using similar reasoning, we can show that REASON entails that C lacks a reason to put his quarter into the vending machine if we take @ to be any of the other worlds in  $H_V$ .

## Wasteful Contributions

There is one case I have not yet considered: the variant of DROPS OF WATER in which there are 15,000 pint holders, the water cart still cannot take more than 10,000 pints, and 10,000 pint holders have already donated their pints. Here, it seems I have no reason to pour my pint into the cart. Doing so would simply be a waste of resources. As Nefsky observes, where you have an outcome-related reason to act, “it needs to be that it is *up in the air* whether or not the outcome in question will occur” (Nefsky 2017: 2758).

The revised version of FULL incorrectly entails that I have an imperfect duty to pour my pint into the cart in this case. By contrast, REASON gives the correct verdict. For when you arrive with your pint, there is no longer any possibility that the men’s suffering will not be alleviated.

This point requires further elucidation. Let us say that it is “up in the air” whether O or O\* will occur, where O and O\* are two incompatible outcomes, when O occurs in at least one world within the deliberatively relevant possibility horizon  $H(t)$  and O\* also occurs in at least one world within  $H(t)$ . Then, REASON entails the *up-in-the-air* condition:

THE UP-IN-THE-AIR CONDITION: You *only* have an outcome-related reason to  $\phi$  rather than  $\psi$  at  $t$  when there are two incompatible outcomes O and O\*, such that it is up in the air whether O or O\* will occur.

To see that REASON entails this, note that there are two ways in which it would not be “up in the air” whether O or O\* will occur: it may be the case that O does not occur in any world within H, or it may be the case that O\* does not occur in any world within H. If O does not occur in any world within  $H(t)$ , O has *infinite negative security* in every world within H. This entails that O is *just as secure* at  $t$  in the world(s) within H where you  $\phi$  at  $t$  as it is in the world(s) where you  $\psi$  at  $t$ . If this is the case, condition (d) is not satisfied. Similarly, if O\* does not occur in any world within H, O\* has *infinite negative security* in every world within H. This entails that O\* is *just as secure* at  $t$  in the world(s) within H where you  $\phi$  at  $t$  as it is in the world(s) where you  $\psi$  at  $t$ . And if this is the case, likewise, condition (d) is not

satisfied. This shows that REASON is *only* satisfied when it is up in the air whether O or O\* will occur.

When you arrive with your pint in our modified version of DROPS OF WATER, the water cart is already completely full. If you pour in your pint, a pint of water will overflow and be wasted. Here, we may assume that there is only one outcome within the possibility horizon: because the cart is already full, the men in the desert will receive a full pint each, whether or not you add your water.<sup>7</sup> Given this, REASON straightforwardly implies that you do not have an outcome-related reason to add your pint. It is not up in the air whether the suffering will be alleviated or not.

## Conclusion

This concludes my discussion of what Nefsky would call ways of rejecting the implication of the inefficacy argument; that is, ways of rejecting premise (i) of this argument. I have argued that non-causal responses to this argument run into either the disconnect problem or the superfluity problem – something which indicates that we need a causal response to this argument. Previous causal responses (Braham & Van Hees 2012; Nefsky 2017; Eriksson 2019) have their own difficulties, but fortunately other causal responses are available. On my preferred account, roughly, you have an outcome-related reason to  $\varphi$  in collective impact cases if and only if your  $\varphi$ -ing increases the security of this outcome. This idea is more precisely formulated in REASON. If REASON accurately identifies the outcome-related reasons we have, (i) is false. You might have an outcome-related reason to  $\varphi$  even if this outcome will occur whether you  $\varphi$  or not. Moreover, we have seen that REASON does seem to accurately identify the outcome-related reasons we have. It gives intuitively correct verdicts in collective impact cases as well as in a broad range of other cases. I will now turn to consider ways of rejecting the description; that is, ways of rejecting premise (ii) of the inefficacy argument.

---

<sup>7</sup> I am assuming here that it is not a live possibility that the cart might somehow be prevented from reaching the men in the desert. If we relax this assumption, the *up-in-the-air* condition does not yield a verdict on the case. However, REASON still delivers the verdict that you do not have a reason to add your pint: because the threat comes from elsewhere – say, from dangers on the road – adding your pint so that a pint of water overflows does not make the good outcome any more secure, or the bad outcome any less secure.

## 7. Denying the Description I

According to the inefficacy argument there are collective impact cases in which the outcome will occur whether or not you act in the relevant way, and because of this you have no collective-outcome-related reasons to act in this way. The conclusion of this argument seems to be false. For it seems that you might have collective-outcome-related reasons to act in the relevant way. So, we need to scrutinise this argument further. There are two ways in which it can be resisted. One is to deny the implication, or in other words assert that you might have a collective-outcome-related reason to act in a particular way *even if* the outcome will occur whether you act in the relevant way or not. The other is to deny the description, and thus hold that the occurrence of the outcome depends on whether you act in the relevant way or not, or at least that this might be the case.

I have considered the first possibility, and argued that you might indeed have a collective-outcome-related reason to act in a particular way *even if* the outcome will occur whether you act in the relevant way or not. In particular, you might have such a reason if your act makes the occurrence of this outcome more secure. In the next three chapters (Chapters 7 through 9), I will consider the second possibility: that the inefficacy argument fails because it is untrue that the outcome will occur whether you act in the relevant way or not, or at least that there is some chance that this is the case.

Philosophical discussion of this issue has centred primarily on how, when and whether the expected utility approach can explain why you have collective-outcome-related reasons. According to the standard view, expected utility can quite straightforwardly explain why you sometimes have such reasons in *threshold cases*, that is cases where there is a threshold such that if  $n$  acts or more of a certain kind are performed some normatively relevant outcome will occur, but if only  $n - 1$  acts or less of this kind are performed this outcome will not occur. In such cases, the occurrence of the outcome might depend on whether you act in the relevant way or not. In THE LAKE,<sup>1</sup> for instance, an expected utility theorist would say that each boat owner has a reason not to use the hazardous paint because there is a risk that life in the lake will come to an end if he, or she, proceeds to use the paint. It is a matter of choice under uncertainty. Still, also according to the standard view, the expected utility approach cannot straightforwardly explain why you might have outcome-

---

<sup>1</sup> This case is introduced on p. 66.

related reasons in *non-threshold* cases, that is cases where no single act makes a difference for whether some normatively relevant outcome occurs or not. To see the problem, consider again DROPS OF WATER. In this case, there are 10,000 people suffering from thirst in a nearby desert, and you are one of 10,000 pint holders each of whom could pour a pint of water into a cart that will be driven to the desert, where the water will then be evenly distributed to those suffering from thirst. If you donate your pint, each person suffering from thirst will get one drop more than they otherwise would have done. If we accept (A) that a single drop of water never can make a perceptible difference to a person's thirst, and (B) that only perceptible differences can matter morally, the expected utility approach entails that you do not have an alleviation-of-suffering-related reason to add your pint of water to the cart. There is no chance that doing so will make a morally relevant difference. More generally, if (A) no act of  $\varphi$ -ing makes a perceptible difference to anyone's harm, and (B) only perceptible differences can matter morally, the expected utility approach says you do not have an outcome-related reason to  $\varphi$ .<sup>2</sup>

Although the discussion of non-threshold cases often revolves around the expected utility approach, non-threshold cases also pose a problem for other theories purporting to explain how you should act on a particular occasion. In particular, they pose a problem for any account of reasons which, like the expected utility approach, relies on SIMPLE.<sup>3</sup> Take objective consequentialism, for instance. That is, consider the principle that you ought always to do what will in fact have the best consequences, where consequences are understood along the lines of SIMPLE (i.e., an event is a consequence of what you did if and only if it would not have occurred if you had acted differently). According to this principle, you lack a reason to pour your pint into the cart in DROPS OF WATER, at least if (A) and (B) hold. Adding a pint will not make things better for anyone. But this seems incorrect. It seems that you do have a reason to pour your pint into the cart. So, one or the other of the assumptions we have made must be mistaken. If you want to hold on to objective consequentialism, you must show that either (A) or (B) is false. The discussion presented in the upcoming chapters is framed around expected utility, but most of what I say would apply to any account of reasons that relies on SIMPLE.

Whether any particular real-life collective impact case is a threshold case or not will depend on the details. Consider climate change, for example. Climate change does involve thresholds of a sort – at some point, as the sea level rises, a house that is situated on the sea front will be flooded; at some point as temperatures rise and malaria spreads to new areas, a certain individual will be infected by this disease, and so on. However, the question is whether these thresholds are fine-grained enough. Could a single drive in a fossil fuel powered car make any difference to

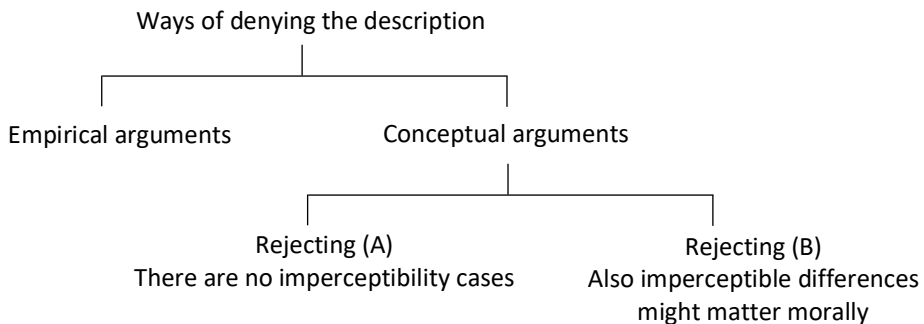
---

<sup>2</sup> In Chapter 8, I call (B) THE PERCEPTIBILITY PRINCIPLE.

<sup>3</sup> This principle is introduced on p. 76.

whether a particular house is flooded, or to whether a particular individual is infected by malaria?

Some have argued on empirical grounds that climate change is a threshold case, much as THE LAKE is (e.g. Broome 2012; Lawford-Smith 2016; Broome 2019). Others maintain that non-threshold cases are impossible. If they are right, there is no need to consider the empirical facts of any specific case in order to know that the expected utility approach applies also to this case. Some of those who argue that non-threshold cases are impossible (e.g. Arntzenius & McCarthy 1997; Norcross 1997, 2004; Kagan 2011) reject (A). They hold it to be a conceptual truth that some act triggers a perceptible difference in harm in all collective harm cases. They would say, for instance, that one or other drop of water necessarily makes a perceptible difference to the sufferers' thirst in DROPS OF WATER. Others (e.g. Parfit 1984; Shrader-Frechette 1987; Barnett 2018; Broome 2019) reject (B). They argue that imperceptible differences in harm can also matter morally. For example, they would say that receiving an additional drop of water when you are suffering from thirst might make a morally relevant difference to the harm you are experiencing even if this extra drop makes no perceptible difference to the way you feel.



In this chapter, I discuss the empirical arguments against the inefficacy argument. In the two following chapters, I discuss the conceptual arguments. In Chapter 8, I consider whether it is true that acts trigger perceptible differences in harm in all collective harm cases. In Chapter 9, which also is the last chapter of this part of the thesis, I consider the remaining questions: Can imperceptible effects be morally relevant? Does the expected utility approach really explain why we have reasons to act in a certain way in triggering cases?

I will argue that appealing to empirical evidence is fine as long as we can trust the expected utility approach, but that the mere possibility of non-threshold cases



should make us hesitate to accept this approach; that we still lack a convincing argument showing that acts trigger perceptible differences in harm in all collective harm cases; that there is, however, a convincing argument showing that harms can be imperceptible (or, alternatively, showing that a set of imperceptible effects might be morally relevant), implying that acting in the relevant way in non-threshold cases might make a morally relevant difference after all, which in turn means that these cases pose no threat to the expected utility approach. Finally, I will argue that the expected utility approach faces another problem. It does not accurately capture all our intuitions about reasons in threshold cases. This becomes particularly clear when we consider ex-post judgements in cases of overdetermination.

## Appealing to Empirical Evidence

Assuming that the expected utility approach gives the right verdict in triggering cases, one could, in some cases, resist the inefficacy argument by arguing that the empirical evidence indicates that doing your part in a collective harm case does make a perceptible difference to the harm. This would amount to arguing that the case at hand is not a collective harm case. Broome (2019), for instance, takes the empirical evidence to indicate that a single drive does make a difference to climate change and its related harms.

Given the atmosphere's instability, we should expect global weather in a few decades' time to be entirely different if you go joyguzzling on Sunday from what it would have been had you stayed at home. [...] It will cause typhoons to form at quite different times and places, and it will lead to a completely different distribution of cholera outbreaks. Your Sunday drive will cause a completely different group of people to be exposed to cholera and other risks of death. Some who would have died will survive because of your drive, and others who would have survived will die. The total numbers who die, and the total amount of harm done in the world may also be greatly altered.

(Broome 2019: 113)

If these empirical details are correct, the expected utility approach entails that you have a reason to refrain from going for a single drive in your car.<sup>4</sup> Some drives make

---

<sup>4</sup> Is Broome right about these empirical details? He builds his case on studies of the butterfly effect, particularly on Palmer, Döring, and Seregin (2014). I contacted the Swedish national weather service, SMHI (the Swedish Meteorological and Hydrological Institute) and asked whether you can draw the conclusions quoted in the main text from the theoretical papers Broome discusses. They answered that, as they define it, "climate" is (roughly) statistics on the weather over a certain period of time, usually several decades, and therefore climate is not directly impacted by the butterfly effect. That is, on the SMHI definition, climate truly is an emergent phenomenon.

a difference for the worse, and others make a difference for the better. Because of the complexity of the situation, we cannot know which difference any particular drive will make. Still, *on average*, single drives makes a difference for the worse, as evidenced by the fact that when enough people drive cars, there is clearly a change for the worse. The expected utility of a single drive is negative.

Alternatively, it can be argued that in any particular case there will be empirical evidence indicating that while most acts make no difference at all, there is some triggering act such that if  $n$  acts of this kind are performed the relevant harm will not occur, and that if  $n + 1$  such acts are performed the relevant harm will occur. This is to argue that the case is a threshold case, much like voting. Lawford-Smith (2016) suggests that climate change fits this bill. Building on climate research, she suggests that there are both macro-level and micro-level thresholds. If a macro threshold is passed, the speed of climate change will increase drastically, and severe harms will ensue. For instance, if the ice sheet covering Greenland were to melt, or the Gulf stream to cease, severe consequences would result. As Lawford-Smith sees it, there are precise thresholds at which these harms are triggered.<sup>5</sup> Micro thresholds are less severe but more common. As an example, she suggests that refraining from using the car one day might “delay a tornado by an hour or a day, or set it on a slightly different path, or cause it to occur with a slightly lesser intensity, all of which would cause slightly less damage to persons than otherwise” (Lawford-Smith 2016: 71). Appealing to the expected utility approach, Lawford-Smith then argues that since we have no way of knowing whether a particular greenhouse gas emitting act will take us past some macro or micro threshold, we have reasons not to perform such acts.<sup>6</sup>

The appeal to empirical evidence rests on two assumptions:

1. THE EMPIRICAL PREMISE: Empirical evidence shows that a single drive in a fossil fuel powered car on average makes a counterfactual difference for the worse to climate change and its related harms.
  2. THE EXPECTED UTILITY APPROACH: If an act on average makes a difference for the worse, you have a reason not to perform it.
- ∴ So, you have a reason not to go for a drive in a fossil fuel powered car.

---

That said, Broome may be right to say that a Sunday drive makes a counterfactual difference to climate-change-related harm in some important sense. The SMHI definition might be too coarse grained to capture all the details of reality.

<sup>5</sup> I think we could question this claim. Still, I will not pursue this line of inquiry here. My focus lies elsewhere.

<sup>6</sup> In his book *Climate Matters*, 2012, Broome argues for a similar view.

If Broome, Lawford-Smith and others are correct, the empirical premise is true. Still, if there are non-threshold cases, the expected utility approach itself is questionable, since non-threshold cases are outright counterexamples to it. It seems that you have an outcome-related reason not to flip the switch in HARMLESS TORTURERS, or to pour your pint into the cart in DROPS OF WATER, but the expected utility approach indicates the contrary. Unless we can show that there are no genuine non-threshold cases, we cannot rely on the appeal to empirical evidence.

To see the threat the mere possibility that there are non-threshold poses to the appeal to empirical evidence, consider what would have happened if the empirical evidence had indicated differently. Suppose that empirical research had revealed that there is no possibility that a Sunday drive in a fossil fuel powered car would make a difference to, or affect, who is exposed to cholera or other risks of harm. Although the atmosphere is unstable, it is not that unstable. And even if there are macro and micro thresholds such that if these thresholds are passed some climate-change-related harm will occur, the emissions from a single drive are incapable of triggering these harms. It is only the accumulation of emissions originating in many drives that could trigger them. It might have turned out that global warming is an emergent phenomenon. Just as one oil molecule never makes a difference to whether or not a small patch of oil is slimy, it might have turned out that one molecule of CO<sub>2</sub> never makes a difference to average temperatures on earth. It might even have turned out that the quantity of molecules released from any single drive makes no difference to average temperatures on earth.<sup>7</sup> If that were the case, the expected utility approach – likewise, other accounts incorporating SIMPLE – would imply that I have no climate-change-related reason to refrain from joy-guzzling. Such a drive has no chance of making a difference to the harm under consideration.

As we can see, if non-threshold cases are possible, the expected utility approach entails that whether or not I have a reason to decline a joy-guzzle depends on physical details of the way in which climate change works at the level of atoms. This verdict, however, seems implausible. Surely, my reason could not depend on such details. If it had turned out that global warming, climate change and their related harms are emerging events, I would still have had a climate-change-related reason not to joy-guzzle.

The same point can be made in relation to HARMLESS TORTURERS. Suppose experiments show that it is true for most victims subjected to the treatment described in HARMLESS TORTURERS that the flipping of some switch triggers pain, but also that there are some victims with a rare neural configuration who are unable to perceive a difference between any adjacent pain states. Would this mean that the torturers have a reason not to flip the switch when a victim with normal neural

---

<sup>7</sup> Kingston and Sinnott-Armstrong (2018) argue that this is the case. I challenge their reasoning in (Gunnemyr 2019).

configuration is connected to the torturing machine but lack that reason when a victim with the rarer condition is connected up? This would be an odd view to hold. Surely, the torturers have a reason not to flip their switches even if the victim is unable to perceive a difference between any adjacent pain states in virtue of the fact the victim will be in excruciating pain when enough switches are flipped.

So, the mere possibility that there are cases in which no act makes a difference to the outcome poses a problem for the expected utility approach, and for any approach that require an act to make, or risk making, a difference to harm's occurrence in order for us to have an outcome-related reason to refrain from performing it. Unless we can show that such cases do not exist, we have reasons to doubt such approaches. This brings us to the question whether non-threshold cases are possible. This is the topic of the next two chapters.



## 8. Making a Vague Difference

Kagan, Nefsky and the Sorites Paradox<sup>1</sup>

(Paper 2)

**Abstract.** In collective harm cases, bad consequences follow if enough people act in a certain way even though no such individual act makes a difference for the worse. Global warming, overfishing and Parfit's (1984) famous case of the harmless torturers are some examples of such harm. Kagan (2011) argues that there is a threshold such that one single act might trigger harm in all collective harm cases. Nefsky (2012) points to serious shortcomings in Kagan's argument, but does not show that his conclusion is incorrect. I argue that our best theories of vagueness (the epistemic view of vagueness, three-valued logic, and supervaluationism) entail that there is a threshold in all collective harm cases. However, my analysis points to another problem with Kagan's argument: the thresholds are not necessarily perceptible. Given the assumption that only perceptible differences matter morally, passing such a threshold does not necessarily trigger morally relevant harm, *pace* Kagan. Last, I consider two variants of Kagan's argument and find both problematic. One controversially assumes that observational relations like "cannot perceive the difference between" are transitive. The other problematically assumes that so called triangulation always is possible.

---

<sup>1</sup> Forthcoming in *Inquiry*.

Collective harm cases arise when bad consequences follow if enough people act in a certain way, even though no such act makes a difference for the worse. For instance, when enough people drive cars running on fossil fuel, this leads to climate change. Still, no single drive makes climate change worse. No extra floods, droughts or storms will occur as a result of one drive. Similarly, when enough people buy factory-farmed chicken, more chickens will be hatched, raised and slaughtered under current factory-farm conditions. However, no future chicken will suffer just because one more grilled chicken is sold.

Since no single act makes a difference for the worse in collective harm cases, it might seem that you have no outcome-related reason to refrain from acting in the problematic way. For instance, since no future chicken will suffer just because you buy one chicken at the supermarket, it might seem that you have no outcome-related reason based on the future suffering of chickens to refrain from buying a chicken.

Then again, you might think that this line of reasoning is mistaken. You might think that you do have outcome-related reasons to refrain from buying factory-farmed chickens, or that you do have outcome-related reasons based on the need to mitigate climate change not to go for a leisure drive. In “Do I make a difference”, Shelly Kagan (2011) argues that there are indeed such reasons, since there is a *risk* that your act will make a difference for the worse. It is a matter of expected utility.

Kagan first identifies two kinds of collective harm case: *triggering* cases and *imperceptible difference* cases. Triggering cases arise when “it is indeed true for most acts that it makes no difference whether or not I do it, but for some act – the triggering act – it makes all the difference in the world” (Kagan 2011: 119). Imperceptible difference cases, on the other hand, are those where no individual act makes a perceptible difference to the amount of harm done. Kagan then argues that the expected utility approach entails that you have an outcome-related reason to refrain from performing the problematic act (e.g. buying factory-farmed chicken) in triggering cases, but not in imperceptible difference cases. He then uses a sorites-style argument to show that – contrary to appearances – it is a conceptual truth that there are no imperceptible difference cases. Instead, all collective harm cases are triggering cases. If this is correct, the expected utility approach can explain why you have reasons to refrain from performing the problematic act in all collective harm cases.

Julia Nefsky (2012) exposes several weaknesses in Kagan’s sorites argument. Importantly, she argues that it amounts to little more than the assertion that plucking a single strand of hair might make someone go bald, and that Kagan’s argument therefore fails to establish that all collective harm cases are triggering cases. Still, although Nefsky finds flaws in Kagan’s argument, she does not show that his conclusion is incorrect. This brings us to a stand-off in the debate.

There have been various reactions to the debate between Kagan and Nefsky. For instance, having set out the dispute, Holly Lawford-Smith (2016) develops an

account which, like Kagan's, crucially depends on there being small enough thresholds. As she sees it, the emission of greenhouse gases from a single flight could trigger the occurrence of some climate-change-related harm. In his book on consumer ethics David T. Schwartz (2017) more or less uncritically endorses both Kagan's position and his argument for it. By contrast, Kai Spiekermann (2014), Zach Barnett (2018), Ewan Kingston and Walter Sinnott-Armstrong (2018), and John Broome (2019) are critical of Kagan's argument.

In this paper, I try to advance the discussion by considering whether our best theories of vagueness – the epistemic view of vagueness, three-valued logic and supervaluationism – entail that imperceptible difference cases are impossible. It turns out that each of these theories entails that there are sharp thresholds in imperceptible difference cases, and this means that Nefsky's objection to Kagan is mistaken (according to these theories). Still, even if our best theories of vagueness entail that there are sharp thresholds in imperceptible difference cases, these thresholds are merely conceptual and therefore not necessarily perceptible. This should not be surprising. We should hesitate before drawing conclusions about perceptibility from truths about language. Moreover, given that only perceptible differences are morally relevant, the thresholds we find in our concepts are not necessarily morally relevant. As a result, we still lack evidence that the expected utility approach works in all collective harm cases.

The details of Kagan's argument are not fully disambiguated. Noting this, I consider two further variants of Kagan's argument. I argue that one controversially assumes that the relation "cannot perceive the difference between" is transitive, and that the other problematically appeals to what Warren Quinn (1990) calls "triangulation".

I proceed as follows. In Part I, I set out the debate between Kagan and Nefsky more fully. In Part II, I consider what the epistemic view of vagueness, three-valued logic and supervaluationism say about Kagan's argument. In Part III, I consider the two variants of Kagan's argument.

## Part I: The Kagan–Nefsky Debate

To illustrate triggering cases, Kagan presents the following example:

**FACTORY-FARMED CHICKEN:** The butcher at the local supermarket will order an additional batch of 25 freshly slaughtered chickens from the local factory farm as soon as 25 chickens have been sold. If he places the order, 25 chickens will be hatched, raised and slaughtered.

Suppose I am considering buying a chicken. Although my purchase has a 24 out of 25 chance of making no difference to the suffering of chickens at all, it has a 1 out



of 25 chance of triggering the suffering of another 25 chickens. Provided that the pleasure I get from eating one chicken does not outweigh the suffering of a chicken raised, hatched and slaughtered under current factory farm conditions, the expected utility of my purchase is negative.

Kagan admits that this example is simplified (and my presentation is even more simplified). Still, it helps us to see why the expected utility approach entails that each of us has a reason not to act in the problematic way in triggering cases. Each of us has such a reason since acting in this way *risks* bringing about a bad outcome. It is a familiar case of a choice under uncertainty.

As presented, the argument may seem to be suggesting that I will make a difference to the suffering of another 25 chickens if and only if I buy the 25<sup>th</sup> chicken. As Kagan clarifies, this is a misinterpretation. For one thing, if 26 chickens are sold that day, my purchase of the 25<sup>th</sup> chicken will have made no difference to the harmful outcome. Another batch would have been ordered anyway, whether or not I had bought a chicken. For another, if my purchase was the last that day, I would not be the only person affecting the harmful outcome. For, each of those who bought one of the first 24 chickens will also have made a difference. It is true of each of them that, had they not bought a chicken, no additional batch of freshly slaughtered chickens would have been ordered. So, rather than making a difference to the harmful outcome if and only if you buy the 25<sup>th</sup> chicken, you make a difference if and only if you buy a chicken and *exactly 25 chickens are sold that day* (or any multiple of 25, such as 50, 75 etc.). Still, buying one chicken has 1 out of 25 chance of triggering the suffering of another 25 chickens.<sup>1</sup>

Unlike triggering cases, imperceptible difference cases arise where many acts together result in a harmful outcome, but where no individual act makes a perceptible difference to the harm. To illustrate such cases, Kagan uses a slightly modified version of Derek Parfit's (1984) case of the harmless torturers.<sup>2</sup>

HARMLESS TORTURERS: A victim is wired to a machine with a thousand switches. There are a thousand torturers, each flipping one of the switches. When no switches are flipped, the victim is in no pain. When all switches are flipped, a strong current runs through the body of the victim, who then is in extreme pain. Still, the flipping of any given switch increases the current only by a very small amount, well below the perceptibility threshold.

---

<sup>1</sup> There are alternative ways of answering this challenge. We could for instance treat at least some triggering cases as examples of pre-emption rather than overdetermination. Lawford-Smith (2016) uses this strategy.

<sup>2</sup> In Parfit's version there are a thousand victims as well as a thousand torturers, and the victims are already in mild pain at the start of the day when no switches are flipped. Kagan's version is presented here in abbreviated form.

Note that in Kagan's version of the case, the victim is in no pain when no switches are flipped. In this respect, Kagan's version differs from Parfit's (and from the version I have discussed earlier in the thesis), where the victim is in mild pain before any switch is flipped. As will become evident, it is important for Kagan's argument that the victim is in no pain at the start of the day. Still, with some ingenuity, his argument could be revised to apply also to Parfit's version of the case.

Kagan suggests that overfishing might also seem to be an imperceptible difference case. If enough people fish, a certain fish stock will collapse, and those who depend on fishing from this fish stock for their livelihood will suffer. Still, no one fish taken makes a difference for whether the fish stock collapses, and so makes no difference for the livelihood of anyone.

As Kagan notes, it is not entirely correct to say that the potentially problematic acts in imperceptible difference cases make no difference at all. They make no perceptible difference to the harm done, but they do make a difference in some underlying dimension. A flipping of a switch increases the current running through the victim, one fish taken means one fish fewer left in the lake, etc. The problem posed by imperceptible difference cases is rather that none of these acts makes a *morally relevant* difference: they do not make any difference in perceived harm for anyone. Here, Kagan presupposes what we can call:

THE PERCEPTIBILITY PRINCIPLE: Only perceptible differences in harm are morally relevant.

This principle has considerable intuitive appeal. How could consequences that are not experienced by anyone – not now, not later – be morally relevant? Moreover, as Kagan points out,

the only harm that seems relevant [in HARMLESS TORTURERS]... is the *pain* the victim is in, and it is difficult to see why it should be bad to increase pain in an imperceptible way. Indeed, it isn't even clear that it makes any *sense* to say that pain has been increased imperceptibly. On the contrary, it seems that the *pain* hasn't increased at all.

(Kagan 2011: 116)

Still, reflecting on examples like this, philosophers like Parfit (1984), Barnett (2018) and Broome (2019) have embraced the initially unappealing position that some harms are imperceptible. In this paper, I will assume with Kagan (2011) that THE PERCEPTIBILITY PRINCIPLE (or something like it)<sup>3</sup> is correct.

---

<sup>3</sup> Shrader-Frechette (1987) argues that THE PERCEPTIBILITY PRINCIPLE is incorrect: imperceptible but measurable differences do matter. As an alternative to THE PERCEPTIBILITY PRINCIPLE, we could

Given that THE PERCEPTIBILITY PRINCIPLE is correct, there is no risk that acting in the problematic way in imperceptible difference cases would make a morally relevant difference. Since no flipping of a switch in HARMLESS TORTURERS makes a morally relevant difference, for instance, you can safely flip a switch knowing that doing so has no risk of making such a difference. Therefore, the expected utility approach does not entail that you have a reason to refrain from acting in the problematic way in such cases.

### **Kagan's Argument and Nefsky's Refutation**

So, the expected utility approach can explain why we should refrain from acting in the problematic way (buying chickens, etc.) in triggering cases, but it cannot explain why we have a reason to refrain from acting in imperceptible difference cases. It is at this point that Kagan's main thesis enters the picture. He argues that it is a conceptual truth that there are no imperceptible difference cases. In his words, "the relevant kind of case is impossible, as a matter of conceptual necessity" (Kagan 2011: 139). If correct, this ensures that all collective harm cases are triggering cases. His diagnosis is that cases that seemed to be imperceptible difference cases are actually triggering cases in disguise. To get the gist of this idea, return to the example of overfishing. Here, Kagan claims that at some point one fish taken does in fact make the fish stock unable to successfully reproduce and thereby cause those who depend on fishing from this fish stock for their livelihood to suffer.

[I]t is not as though my catching an extra fish makes an imperceptible difference to the decline of the fish population's ability to replenish itself. Rather, up to a point, as each person takes one extra fish, this makes no difference at all – the fish will be fully capable of replenishing their population, despite the extra fish taken. But at a certain point – an unknown point – one too many fish is taken, the stock is no longer large enough, and the population crashes.

(Kagan 2011: 118)

Similarly, Kagan argues, the flipping of a switch in HARMLESS TORTURERS makes a perceptible difference in pain for the victim.

Kagan has offered an argument for the view that all collective harm cases are triggering cases, using HARMLESS TORTURERS as an illustration. However, though

---

adopt a measurability principle saying that an action cannot be right or wrong as a result of its effects if those effects are immeasurable. Since infinitesimal changes are not measurable, there could be cases of immeasurable difference. See Nefsky (2012: sec. XII) and Broome (2019: appendix) for discussion. Kagan does not consider a measurability principle.

the general idea behind the argument is clear, it is stated somewhat differently in different places in Kagan's paper. I will therefore consider three variants of it.

I start with Nefsky's (2012) interpretation. In summary, she understands Kagan's argument as follows:

ARGUMENT #1 (REPORT VERSION)

- (1) When no switches are flipped the victim is in no pain (call this  $s_0$ ), and if asked whether he is in pain, he will answer "No".<sup>4</sup>
- (2) Suppose (for *reductio*) that, (a) for all  $n$ , the victim cannot perceive the difference between  $s_n$  and  $s_{n+1}$ .<sup>5</sup> If this is true, (b) the victim (who always gives a correct report of what he feels) will give the same answer to the question "Are you in pain?" in  $s_{n+1}$  as in  $s_n$ .
- (3) Therefore, when in  $s_{1000}$ , if asked whether he is in pain, the victim will answer "No".

This is absurd! By hypothesis, the victim is in extreme pain when in  $s_{1000}$  and will therefore answer "Yes" when asked whether he is in pain.

∴ Supposition (2a) must be wrong. There must be some  $n$  such that the victim can perceive the difference between  $s_n$  and  $s_{n+1}$ .<sup>6</sup>

This argument, Kagan claims, can be repeated for any alleged imperceptible difference case.

As Nefsky (2012) points out, argument #1 (report version) is not compelling. It relies on the assumption that the victim always gives a correct report of what he feels (an assumption made explicit in (2b)). This assumption is however highly questionable. The requirement that the victim must answer with a simple "Yes" or "No" makes it unlikely that his report will reflect his experiences. His pain increases gradually in a fine-grained manner, but he is forced to give coarse-grained answers.

---

<sup>4</sup>  $S_n$  denotes pain state  $n$ , that is, the pain the victim feels when  $n$  switches are flipped.

<sup>5</sup> Towards the end of his paper, Kagan argues that it would not be genuinely troublesome for the consequentialist that the victim cannot perceive the difference between two adjacent pain states. What *is* genuinely troublesome, Kagan suggests, is "the suggestion that an act might leave someone worse off, even though it makes no difference to how they feel" (2011: 137). This seems correct. Arguably, it is not the victim's ability to distinguish two adjacent states that is morally relevant, but rather differences in how the victim feels. If you wish, you are free to replace (2a) with the following alternative premise: for all  $n$ , the victim feels the same in  $s_{n+1}$  as in  $s_n$  (and to apply this premise throughout the paper, including to all versions of Kagan's argument). My arguments stand either way.

<sup>6</sup> Norcross (1997: 142) suggests a similar argument in which the victim screams when he is in pain rather than reporting "Yes".

To avoid this difficulty, Nefsky recasts Kagan's argument without references to what the victim reports.

#### ARGUMENT #1

- (1) When no switches are flipped the victim is not in pain (call this  $s_0$ ).
- (2) Suppose (for *reductio*) that, (a) for all  $n$ , the victim cannot perceive the difference between  $s_n$  and  $s_{n+1}$ . Therefore, (b) for all  $n$ , if the victim is in no pain in  $s_n$ , he is also in no pain in  $s_{n+1}$ .
- (3) Therefore, when in  $s_{1000}$ , the victim is in no pain.

This is absurd! By hypothesis, the victim is in extreme pain when in  $s_{1000}$ .

∴ Supposition (2a) must be wrong. There must be some  $n$  such that the victim can perceive the difference between  $s_n$  and  $s_{n+1}$ .<sup>7</sup>

Nefsky points to a flaw also in this improved version of the argument. We are not compelled to conclude that (2a) is wrong. In fact, we are not compelled to locate the problem in (2) at all. Rather, we face a dilemma. We need to reject *either* at least one premise *or* the validity of the argument, but neither option seems attractive. In Nefsky's words, this is an argument where "apparently true premises are shown to lead to contradiction through a series of seemingly valid steps" (Nefsky 2012: 385, footnote 48). If argument #1 can be used to show that the victim can perceive the difference between one pain state and the next, we can use an analogue of it to show that adding a grain of sand can turn a non-heap into a heap, or that plucking a single strand of hair from a hairy person can make this person bald. As Nefsky puts it, Kagan is basically "giving a sorites argument as though it were a simple *reductio* proof that there cannot be vague boundaries" (Nefsky 2012: 385).

To bring out this point clearly, we can compare argument #1 with a generic sorites argument:

#### THE SORITES ARGUMENT

- |   |                     |
|---|---------------------|
| (1) $x_1$ is $F$  | (Base premise)      |
| (2) For all $n$ , if $x_n$ is $F$ then $x_{n+1}$ is $F$ | (Inductive premise) |
| (3) Therefore, $x_k$ is $F$                             | (Conclusion)        |

Paradox! Even though both (1) and (2) appear to be true, and (3) appears to follow from (1) and (2), (3) seems false for sufficiently large  $k$ 's.

---

<sup>7</sup> This is the argument that Nefsky (2012) calls "Version 2". The argument is presented here in condensed form.

Comparing the two arguments, we see that the conclusion (∴) in argument #1 lacks a counterpart in the sorites argument. On reflection, we see that (∴) points to a solution to the paradox, because it suggests that the problem is located in the inductive premise (2). However, just as the sorites argument does not force us to accept, say, that adding a grain of sand can turn a non-heap into a heap, premises (1) to (3) do not force us to reject (2). There is a range of possible solutions to the paradox, all of which have their own problems. We could, for instance, accept conclusion (3). Peter Unger (1979) advocates this. He argues that since (1) and (2) are true, and since the argument is valid, we must draw the conclusion, however unintuitive it may seem, that (3) is true. Adding one grain of sand to a non-heap will never result in a heap. In fact, there are no heaps, hairy persons or extreme pains. Hence the title of his paper: “There are no ordinary things”. If Unger is correct, there will be no sharp boundary in imperceptible difference cases, and Kagan’s argument will be unsound.

Another option is to deny the validity of the argument. Bertrand Russell (1923) claims that arguments containing vague predicates are not the kind of thing that can be valid or invalid. If he is correct, argument #1 is invalid since it contains a vague predicate. A third option is to follow Michael Dummett (1975) and accept (1), (2) and (3) and the validity of the argument, and conclude that the predicate in question is incoherent. If we follow Dummett’s advice, we will not conclude that (2) is wrong as Kagan urges us to do. We will conclude instead that the predicate represented by *F* (e.g. “not a heap” or “pain”) is incoherent. If any of these accounts turns out to be correct, argument #1 will lack cogency.

This brings us to what is essentially a crux of the debate. Nefsky argues that argument #1 is not compelling, and that we are not forced to reject (2). But she does not show that Kagan’s conclusion is mistaken. It might still be a conceptual truth that the victim can perceive the difference between some adjacent pain states in HARMLESS TORTURERS – and, more generally, that imperceptible difference cases are impossible. Nefsky proposes a range of other objections to Kagan’s argument. However, none of these shows that Kagan necessarily is wrong to reject (2). Nefsky does, however, say the following:

[W]e should begin our evaluation of Kagan’s arguments – as he begins his arguments – from the intuitive, pre-theoretic perspective in which nontriggering cases seem to be a real possibility.

(Nefsky 2012: 378, note 35)

From a pre-theoretic perspective, cases like HARMLESS TORTURERS do at least seem to be non-triggering cases (of the imperceptible difference variety). Nefsky also appears to be correct in claiming that we should begin our evaluation from this perspective. But it seems that the natural next step would be to move on from a pre-

theoretic perspective to a theoretical perspective. What do our best theories of vagueness tell us about these issues? Do they tilt the scale in favour of Kagan's position or indicate that his conclusion is mistaken?

## Part II: Theories of Vagueness and Kagan's Argument

To determine whether it is conceptually true that there is a morally relevant boundary in all imperceptible difference cases we might ask what our best theories of vagueness say. I am not thinking of the accounts, already mentioned, presented by Unger, Russell and Dummett. I mentioned those just to show that there are putative solutions to the sorites paradox other than Kagan's rejection of the inductive premise. Instead, I am thinking of the epistemic view of vagueness, three-valued logic and supervaluationism. Do these theories entail that there is a sharp, morally relevant boundary in imperceptible difference cases?<sup>8</sup> The question whether Kagan's argument is sound would no doubt be easier to settle if we all agreed on one theory. However, judgements on this issue differ. I will not pretend to be able to settle the question of which of these theories (if any) is the correct one. Instead, I will consider whether the current main candidates for explaining sorites cases imply that (2) must be rejected.

### **There Are Sharp Boundaries in Vague Concepts, Says Our Best Theories of Vagueness**

Premise (2) consists of two statements. The first, (2a), is alleged to imply the second, (2b). In this section, I want to concentrate on (2b). Does the epistemic view, three-valued logic or supervaluationism entail that (2b) is false: that it is not the case that, for all  $n$ , if the victim is in no pain in  $s_n$ , he is also in no pain in  $s_{n+1}$ ? To anticipate, the answer is yes. In fact, each of the three theories entails that (2b) is false. This means Kagan's argument may be sound after all.

We can begin with the epistemic view of vagueness.

THE EPISTEMIC VIEW: Vagueness is due to ignorance. Vague predicates have unknown sharp boundaries, and therefore predications in borderline cases are either true or false – we just do not know which.

(See Cargile 1969; Sorensen 1988; Williamson 1994)

---

<sup>8</sup> Since Kagan (2011) argues that it is a conceptual truth that imperceptible difference cases are impossible, I focus on theories of vagueness in language. I set theories of ontic vagueness such as that given in Barnes (2010) aside.

According to this view, there is a unique and precise point at which a growing person suddenly becomes tall, at which the addition of a grain of sand turns a non-heap to a heap, and at which the loss of a single hair makes a person bald. Likewise, in HARMLESS TORTURERS, there is a unique and precise threshold beyond which the victim is in pain. We just do not know exactly where this threshold is, and neither does the victim. On this view, we can reject the inductive premise in sorites cases, and this will leave us with room to agree that Kagan is correct.

The epistemic view might seem incredible in its claim that there are sharp boundaries in all sorites cases. Still, it has the theoretical advantage of preserving classical logic – this is why its leading proponents typically put considerable effort into explaining why various nonclassical logics fail. However, instead of retaining classical logic and accepting the epistemic view, we could accept a non-classical logic such as three-valued logic.

THREE-VALUED LOGIC: For borderline cases, it is neither true nor false that  $Fx$ . Instead,  $Fx$  takes the intermediate third truth-value *indefinite*.

(See Halldén 1949; Łukasiewicz 1970/1920; Tye 1990, 1994)<sup>9</sup>

According to this view, it is neither true nor false, but indefinite, whether a person who is (say) 178 cm is tall or not. Likewise, when (say) 120 switches are flipped in HARMLESS TORTURERS, it is indefinite whether the victim is in pain. Typically, proponents of three-valued logic think that vagueness is due primarily, not to ignorance, but semantic imprecision (see e.g. Tye 1990).

Interestingly, according to three-valued logic, there will be a sharp boundary at which a growing person becomes tall, and at which flipping a single switch could make the sentence “the victim is in pain” true. Why? Since there are only three possible truth values, there must be some place in a sorites series when the truth value changes from indefinite to true. It cannot be indefinite whether the sentence “the victim is in severe pain” is true or indefinite.

In fact, according to three-valued logic, there are two sharp boundaries in a sorites case: one between the false and indefinite cases, and one between the indefinite and true cases. For instance, there is a flipping of a switch such that the statement “the victim is in pain” becomes indefinite, not false, and there is a different flipping of a switch such that the statement “the victim is in pain” ceases to be indefinite and becomes true.

---

<sup>9</sup> There are some different suggestions for how to denote and understand the third truth-value. For instance, Tye (1994) denotes it “indefinite”, Halldén (1949) calls it “meaningless” and Łukasiewicz (1970/1920) labels it “possible”.





Adherents of three-valued logic will reject the inductive premise of the sorites argument. It is not true for all  $n$  that if  $x_n$  is  $F$ , then  $x_{n+1}$  is  $F$ . So, three-valued logic implies that Kagan is correct. It is a conceptual truth that the flipping of a switch might make the statement “the victim is in pain” true.<sup>10</sup>

Finally, supervaluationism might turn out to be our best theory of vagueness. Like most proponents of three-valued logic, supervaluationists like Kit Fine (1975), David Lewis (1986c) and Rosanna Keefe (2000) take vagueness to stem from our indecisiveness about how to use our concepts, not our ignorance. In a passage that is often quoted, Lewis writes:

The only intelligible account of vagueness locates it in our thought and language. The reason it’s vague where the outback begins is not that there’s this thing, the outback, with imprecise borders; rather there are many things, with different borders, and nobody has been fool enough to try to enforce a choice of one of them as the official referent of the word “outback”.

(Lewis 1986c: 213)

Presumably, we can understand the vagueness of the statement “the victim is in pain” in a similar way. It is not that there is this thing, the victim’s being in pain, imprecisely described; rather there are many things, each precisely described, and nobody has been fool enough to try to enforce a choice of one of them as the official referent of the “the victim’s being in pain”. As Fine (1975) puts it: “vagueness is ambiguity on a grand and systematic scale” (282).

Supervaluationists then claim that a vague statement like “the victim’s being in pain” is true (in the sense they explain) just when it is true on each acceptable description; that is, on each acceptable way of making it perfectly precise. This idea is commonly phrased in the following way:

SUPERVALUATIONISM: A statement is supertrue and therefore true if and only if it is true on all its acceptable completely sharp sharpenings.

(See Fine 1975; Keefe 2000; Broome 2004)

---

<sup>10</sup> Notice that, in general, many-valued logics (four-, five-valued logics, and so on) will entail that a single flipping of a switch may make the statement “the victim is in pain” true.

For example, the statement “Tek is tall”, said about Tek, who is 178 cm, is not supertrue (and hence not true), because there are completely sharp sharpenings of “tall” according to which 178 cm does not count as tall: “180 cm or taller” could be an acceptable and completely sharp sharpening of “tall”, for instance.<sup>11</sup> Similarly, assuming that it is vague whether the victim is in pain when 120 switches are flipped in HARMLESS TORTURERS, the supervaluationist can hold that it is not true (because not supertrue) that the victim is in pain. There are acceptable sharpenings of “being in pain” according to which the victim is in pain, and acceptable sharpenings of “being in pain” according to which he is not in pain. Things would have been different if Tek had been 198 cm, or if 700 switches had been flipped. There is arguably no acceptable sharpening of “tall” according to which a height of 198 cm is not tall, and no acceptable sharpening of “being in pain” according to which the affective state a torture victim is in when 700 switches are flipped is not pain.

Instead of introducing an additional truth-value (as three-valued logic does), supervaluationism accommodates truth-value gaps. For instance, if Tek has a height of 178 cm, the statement “Tek is tall” is neither true nor false. However, this does not mean that this statement assumes some further, third truth value. Likewise, when 120 switches are flipped in HARMLESS TORTURERS, it is neither true nor false (nor indefinite) that the victim is in pain. For this reason (and a few others), supervaluationism is substantially compatible with classical logic.

According to supervaluationism, then, (2b) is false. On each completely sharp sharpening of “being in pain”, there is a sharp boundary between the pain states where the victim is in pain and the pain states where he is not. This may seem strange. The following examples may help to explain how it can be so. Consider the statement:

(A) There is a sharp boundary between persons who are tall and persons who are not.

However we sharpen “tall”, (A) will be true. This means that (A) is supertrue – and therefore true. If, for instance, we sharpen “tall” so that a person is tall if and only if they are taller than exactly 178 cm, there will be a sharp boundary between people who are tall and people who are not. Everyone will either belong to the set of those who are tall or to the set of those who are not tall. Similarly, if we sharpen “tall” so that a human being is tall if and only if they are taller than exactly 180 cm, there will be a sharp boundary between tall and non-tall people, although of course slightly fewer individuals will now belong to the set of those who are tall.

The same goes for the following statement:

---

<sup>11</sup> In what follows, when I talk about sharpenings, I mean completely sharp sharpenings.

- (B) There is a sharp boundary between the states in which the victim is in pain and the states in which he is not in pain.

This statement will be true on all sharpenings of “in pain”, and therefore supertrue, and therefore true. Generally, statements of the form “There is a sharp boundary between the states in which  $x$  is  $F$  and the states in which  $x$  is not  $F$ ” come out true on all sharpenings of  $F$  (see Fine 1975; Keefe 2000). Putting the point differently, we can say that supervaluationism entails that the inductive premise (For all  $n$ , if  $x_n$  is  $F$  then  $x_{n+1}$  is  $F$ ) is false. On each sharpening of  $F$ , there will be some  $n$  for which  $x_n$  is  $F$  and  $x_{n+1}$  is not  $F$ .<sup>12</sup>

So, the epistemic view, three-valued logic and supervaluationism all entail that (2b) is false. A sharp boundary separates cases where the victim is in pain and cases where he is not. This opens up the possibility that Kagan’s conclusion is correct after all. It might be a conceptual truth that there are no imperceptible difference cases. Our pre-theoretic verdict about the inductive premise is most likely mistaken.<sup>13</sup>

### Another Way to Reach the Same Conclusion

We could reach the same conclusion without the apparatus of the epistemic view, or three-valued logic or supervaluationism. Argument #1 prompts us to reject one of the following:

- (1) when no switches are flipped, the victim is not in pain,
- (2b) for all  $n$ , if the victim is in no pain in  $s_n$ , he is also in no pain in  $s_{n+1}$ ,
- ( $\neg$ 3) the victim is in pain when 1,000 switches are flipped, or
- the validity of the argument.

Kagan thinks we have to reject (2b), and by extension (2a). Nefsky argues that we are not forced to do so, as we might instead reject (1), ( $\neg$ 3) or the validity of the argument. However, Nefsky’s argument is not fully convincing. She is right that we

---

<sup>12</sup> Note, however, that it is indeterminate where the sharp boundary is located. This is indeterminate because the location of the boundary differs from sharpening to sharpening. The range of admissible sharpenings is also indeterminate.

<sup>13</sup> We cannot definitely conclude that our pre-theoretic verdict is mistaken. Other theories of vagueness (e.g. the already mentioned accounts by Unger, Russell and Dummett) locate the problem elsewhere. These accounts, however, have problems of their own (see e.g. Keefe 2000: 18-26). So, we will still do best to rely on the verdicts of the epistemic view, three-valued logic and supervaluationism.

are not forced to reject (2b). However, on closer reflection, it would seem that doing so is likely the best option. It seems clearly wrong to deny (1). Likewise, the idea that the victim is not in excruciating pain when all the switches are flipped is farfetched, to put it mildly, so  $(\neg 3)$  is safe. Moreover, the argument is almost certainly valid. As Dummett (1975) argues, to reject sorites arguments as invalid we must give up some fundamental rules of inference, and doing so would be a high price to pay. Sorites arguments rely merely on the inference rules of *modus ponens* and universal instantiation. The rejection of these rules would have far-reaching effects.<sup>14</sup> This leaves us with (2b). This premise might appear correct at first sight, but it is not as obviously correct as the other premises, or as difficult to challenge as the validity of the argument. So, it appears that were we forced to identify a problem with this sorites argument, we should point to (2b). Given this, we have independent reasons to think that the verdict on this issue given by the epistemic view, three-valued logic and supervaluationism is correct.

### **Conceptual Sharp Boundaries Do Not Have to Be Perceptual**

While each of the epistemic view, three-valued logic and supervaluationism implies that (2b) is false, none of these accounts implies that (2a) is false. This indicates that there is a problem in the implication from (2a) to (2b). To explain how this can be the case, we have to go back and consider the different ways of handling vagueness more closely.

On the epistemic view, there is always an unknown sharp boundary in sorites cases. However, this boundary does not have to correspond to a sharp boundary in the world. It may instead correspond to a sharp boundary in our language. According to Timothy Williamson (1994), for instance, the location of the sharp boundary is determined by the meaning of the vague predicate in question, and meaning, in turn, supervenes on use. If we are to determine whether the victim is in pain when 120 switches are flipped, we have to consider the meaning of the predicate “is in pain”. This in turn requires us to consider how this predicate is used in ordinary language. In many cases, the meaning of a predicate is stabilised by natural divisions, making small changes in use irrelevant. With this in mind, Williamson says, a “slightly increased propensity to mistake fool’s gold for gold would not change the meaning or the extension of the word ‘gold’” (1994: 231). However, he warns that this is not true of vague predicates, since where these are concerned there are no natural divisions that might help stabilise the boundary. When it comes to vague predicates, meaning supervenes wholly on use. So, even if, on the epistemic view, there is a sharp boundary in HARMLESS TORTURERS, this boundary does not exist because the

---

<sup>14</sup> If you think that the universal quantifier might give rise to validity problems, argument #1 could be restated without it, using only *modus ponens*.

victim perceives it. Its existence supervenes on the way we use the predicate “is in pain” in ordinary language.

Elaborating this point, we can note that Williamson (1994: 180-84) also argues that even if an object is red, an observer will not necessarily know that it is red. The observer might be mistaken. This, he continues, goes for any observational property, including “being square” and (presumably) “being in pain”. Consider again HARMLESS TORTURERS, and imagine that we flip the switches one at a time, and that after flipping each switch we ask the victim, “Are you in pain?” It seems likely that at some interval in this series of switch-flipping the victim would be unsure how to answer. Perhaps when the 120<sup>th</sup> switch is flipped, he will answer: “Well, I feel something, but I am unsure of whether I should call it pain.” As more and more switches are flipped, he might become increasingly confident that he has passed the threshold for being in pain, and if we require him to give a simple “Yes” or “No” answer, he will eventually stop answering “No” and say “Yes”. There is nothing in this story to indicate that the victim would necessarily perceive a difference in pain between two adjacent states.

The upshot is that the epistemic view does not ensure that there is a perceptual difference of the kind Kagan needs. The victim might be in pain without perceiving pain (i.e. the statement “the victim is in pain” could be true even where the victim does not recognise what he is feeling as pain), and there might be a difference in pain even though the victim does not perceive any difference in the pain.

Three-valued logic runs into a similar problem, and for similar reasons. While it implies that a sharp conceptual boundary marks the point at which the statement “the victim is in pain” becomes true rather than indefinite (and another where this statement becomes indefinite and no longer false), three-valued logic fails to imply that there is a sharp perceptual boundary where the victim starts being in pain. And so, again, it will not deliver Kagan the result he needs.

Likewise for Supervenience, but for a different reason. One might assume that if a statement about the world is true, this truth will correspond to something in the world. However, if we take supervenience at face value, this is not always the case. As Bertil Rolf (1981) points out, a statement like “there exists a sharp boundary between red and pink” comes out as true even though there is no such boundary. It does so because on each completely sharp sharpening of “red” and “pink” there is a sharp boundary between these colours. Rolf questions supervenience as such on these grounds.

Faced with Rolf’s objection, supervenience has two options: they can bite the bullet and agree that supervenience separates statements of existence from actual existence, or they can restrict the range of statements to which supervenience applies. Keefe (2000) and Lewis (1993) take the first option. Broome (2004) takes the second. Either way, supervenience does not entail that the victim can perceive a difference in pain between a certain pair of adjacent

affective states. If we bite the bullet, the fact that supervenience entails that (B) is true will tell us nothing about what the victim perceives. And if we restrict the range of statements to which supervenience applies in order to avoid biting the bullet, supervenience will not be applicable to (B) (my earlier discussion implicitly assumed that it was). In what follows, for the sake of simplicity I will assume that supervenience is applicable to statements like (B).

According to the supervenience theorist, then, there is sharp boundary between the states in which the victim is in pain and the states in which he is not (i.e. (B) is true), even if there is no sharp boundary separating the states in which the victim perceives pain and the states in which he does not.

To sum up, the epistemic view, three-valued logic and supervenience all entail that there are sharp conceptual boundaries in imperceptible difference cases like HARMLESS TORTURERS. However, they do not entail that these boundaries are perceptible. Where argument #1 is concerned, they entail that (2b) is false even if (2a) is true. This means that the implication from (2a) to (2b) is invalid. In turn, this means that argument #1 does not show what Kagan wants it to show: that, necessarily, there is a perceptible threshold in all collective harm cases. This should not surprise us. Even if it is a conceptual truth that there is a sharp border between the states where the victim is in pain and those where he is not, this tells us nothing about whether the victim can perceive this difference. The first point is about language, and the second is about perception.

Moreover, since argument #1 fails to show that the flipping of a switch makes pain perceptibly worse, it also fails to show that the expected utility approach gives the right verdict in all collective harm cases. Only perceptible differences are morally relevant according to THE PERCEPTIBILITY PRINCIPLE, and because there is still a live possibility that no flipping of a switch makes a perceptible difference in pain, it may still be the case, for all that has been said, that no flipping of a switch makes a morally relevant difference.

### Part III: Two Other Versions of Kagan's Argument

Have we been considering the wrong argument? Instead of considering argument #1 (which relies on the implication from (2a) to (2b) being valid), perhaps we should have examined an argument where something like (2a) on its own constitutes the inductive premise in the argument. In other words, we should concentrate on whether the victim can *perceive* the difference between a certain pair of adjacent pain states rather than on whether there is a pair of adjacent pain states such that we *can say* that he, the victim, is in pain in one but not the other. The alternative argument I have in mind could be spelled out as follows (again, I have ignored the victim's reports):

ARGUMENT #2

- (1) When no switches are flipped the victim is not in pain (call this  $s_0$ ).
- (2) Suppose (for reductio) that, for all  $n$ , the victim cannot perceive the difference between  $s_n$  and  $s_{n+1}$ .
- (3) Therefore, the victim cannot perceive the difference between  $s_{1000}$  and  $s_0$ . So, when in  $s_{1000}$ , the victim is not in pain.

This is absurd! By hypothesis, the victim is in extreme pain when in  $s_{1000}$ .

- ∴ Supposition (2) must be mistaken. There must be some  $n$  such that the victim can perceive the difference between  $s_n$  and  $s_{n+1}$ .

When Kagan first sets out his argument, he presents it along the lines of argument #1. However, at other places, he seems to have this second argument in mind. For instance, when commenting on his argument, he writes:

It simply cannot be that every state feels like the one before it, for by hypothesis state 0 feels like no pain, while state 1,000 feels like pain. Hence at least one state must feel different from the one that came before.

(Kagan 2011: 132)

Setting exegetical issues aside,<sup>15</sup> we might ask whether argument #2 is cogent. The answer is that it is not. While, commendably, it declines to draw conclusions about perception from previously reached conclusions about language, it does assume, questionably, that the relation “cannot perceive the difference between” is transitive. It assumes that we can infer that (3) the victim cannot perceive the difference between the first and the last pain state from the fact that (2) he cannot perceive the difference between any two adjacent pain states.

Standardly, philosophers have taken relations like “looking the same as” and “tasting the same as” to be non-transitive (this goes for, for instance Goodman 1951; Armstrong 1968; Dummett 1975; Tappenden 1993), and for quite a long time and Frank Jackson and R. J. Pinkerton (1973) were alone in questioning the non-transitivity of this kind of relation, at least as far as I know. More recently, however, Diana Raffman (2000) and Fara Delia Graff (2001) have argued for the transitivity of perceptual indiscriminability, arguing that sorites series of the phenomenal kind are impossible, effectively proving Kagan right. Keefe (2011) has defended the non-transitivity of these relations, arguing that there is no good reason to think that we should treat phenomenal sorites differently from non-phenomenal ones. I will not go through the arguments for and against the idea that perceptual indiscriminability

---

<sup>15</sup> Kagan may not have intended his argument to be interpreted in the way canvassed here.

is a transitive relation – that would take us too far astray. For now, it is enough to say that the standard view here is that perceptual indiscriminability is not a transitive relation, so argument #2’s assumption that it is is problematic.

Instead of understanding relations such as “feels the same as” and “cannot perceive the difference between” in terms of vagueness, as Keefe urges us to do, we could understand them in terms of parity. If we do, we again find them to be intransitive. It has been suggested that there is a fourth comparative value relation besides the standard three of “better than”, “worse than” and “equally good”. This is variously called “parity” (Chang 2002), “incommensurability” (Rabinowicz 2009), “rough comparability” (Temkin 2012) and “imprecise equality” (Parfit 1984). The idea is that when two things, A and B, are on a par, A is not better or worse than B, but nor is it the case that A and B are equally good. One significant difference between exact equality and parity is that while the former is transitive, the latter is not. Oasis might be on a par with Blur, and Blur might be on a par with Pulp, even though Pulp are better than Oasis. Just as Britpop bands from the 1990s could be on a par in terms of their value, perhaps pain states could be on a par in terms of how they feel. If that were the case, it could be true that the victim in HARMLESS TORTURERS cannot perceive the difference between a certain pair of adjacent pain states but can perceive a difference between two pain states that are not adjacent, i.e. are more distant from each other.<sup>16</sup>

Here, I do not wish to take a stance on how to best capture the idea that the relation “cannot perceive the difference between” might lack transitivity. Is it because of vagueness? Or because of parity?<sup>17</sup> It is sufficient for my purposes to show only that argument #2 assumes that “cannot perceive the difference between” is a transitive relation, and to note that this is an assumption that needs to be carefully defended.

### **A Third Version of Kagan’s Argument: Triangulation**

When discussing the idea that the victim in HARMLESS TORTURERS is able to perceive the difference between a pair of adjacent pain states, Kagan (2011) concedes that “this difference in pain might not be noticed if the victim limits himself to direct pairwise comparisons between the two adjacent states” (Kagan 2011: 136). Instead, he argues, the difference between two adjacent pain states might become obvious only when the victim compares each of the two pain states to some third baseline state. For instance, while the victim may not be able to

---

<sup>16</sup> Similarly, Spiekermann (2014) argues that non-triggering cases are possible because experiences of harm are intransitive, and Hedden (2020) reasons that some collective harm cases are non-triggering cases because they involve parity.

<sup>17</sup> Philosophers like Broome (2004) and Andersson (2017) have argued that there is no need for a fourth value relation like parity. They say hard cases of comparison are cases of vagueness. Others, like Chang (2002) and Rabinowicz (2009), disagree.



distinguish state  $s_{66}$  from  $s_0$ , he may nevertheless be able to distinguish the next state,  $s_{67}$ , from  $s_0$ . That would put him in a position to infer that  $s_{66}$  and  $s_{67}$  differ – an inference that has been called “triangulation” (e.g. by Quinn 1990). Kagan invokes this idea as a way of making the conclusion that, necessarily, the victim can perceive the difference between some adjacent pain states more palatable. However, we can easily turn the idea into an argument on its own. It would go like this (I will continue to ignore the victim’s reports):

Suppose that (1) the victim cannot perceive the difference between  $s_0$  and  $s_1$ . Suppose also that (2) if the victim cannot perceive the difference between  $s_n$  and some baseline state such as  $s_0$ , then he cannot perceive the difference between the next pain state  $s_{n+1}$  and the baseline state  $s_0$ . Then the victim cannot perceive the difference between  $s_0$  and  $s_2$ , which in turn means that he cannot perceive the difference between  $s_0$  and  $s_3$ , ..., which means that (3) he cannot perceive the difference between  $s_0$  and  $s_{1000}$ . This conclusion is clearly wrong. The victim is in severe pain when a thousand switches are flipped and is definitely able to perceive the difference between this pain state ( $s_{1000}$ ) and the first one ( $s_0$ ), so one or other of the suppositions we have made must be mistaken. Since it seems overwhelmingly likely both that the victim cannot perceive the difference between  $s_0$  and  $s_1$ , and that the victim can perceive the difference between  $s_0$  and  $s_{1000}$ , and that the inference rules are valid, we should reject (2) even if this premise seemed plausible at first sight. That is, there must be some number of switches,  $n$ , conforming to the following specification: the victim cannot perceive the difference between the pain states  $s_n$  and  $s_0$  but can perceive the difference between pain states  $s_{n+1}$  and  $s_0$ . Call this “argument #3”. Similar arguments have previously been put forward by Rabinowicz (1989), Carlson (1996) and Norcross (1997).

Argument #3 avoids the objections I have raised to arguments #1 and #2. It does not assume that we can infer truths about perception from truths about concepts, and it does not assume that the “cannot perceive the difference between” relation is transitive. Still, this argument does not necessarily hold since our perceptions might show higher-order indeterminacy.<sup>18</sup>

---

<sup>18</sup> As it stands, argument #3 does not show that flipping one switch has a 1/1000 chance of triggering excruciating harm. At most, it implies that some flipping of a switch triggers pain that is too small to be directly perceptible, but that is detectable through triangulation. Still, it seems that this gap could be bridged, perhaps roughly in the following way. Suppose that argument #3 holds. Suppose also that experience shows that there is a triangulable difference between  $s_{66}$  and  $s_{67}$ , and that there is still a significant difference between  $s_{67}$  and  $s_{1000}$ . In that case, we could apply argument #3 once more, using  $s_{67}$  as baseline, showing that there is at least one more triangulable difference between  $s_{67}$  and  $s_{1000}$ . We could continue in this manner until we reach a pain state that is indistinguishable from  $s_{1000}$ , showing that there is a certain number of triangulable differences in total. It seems that these triangulable differences in pain must amount to the total pain. If there are only a few triangulable differences in pain between the first and the last pain state, each triangulable difference must involve considerable pain. And in cases like HARMLESS TORTURERS,

For a start, there might be first-order indeterminacy. There might be a range of pain states for which it is indeterminate whether they are different from the baseline or not. It might for instance be unclear to the victim whether  $s_{66}$ ,  $s_{67}$ , ..., and  $s_{70}$  are perceptibly different from the baseline. (Here, you can think about indeterminacy in your own preferred way: either as a third truth-value or as a truth-value gap.) Because of this possibility, the falsity of premise (2) does not necessarily entail that there is a pain state  $s_n$  such that  $s_n$  is not perceptibly different from the baseline, while  $s_{n+1}$  is. It could instead be indeterminate whether  $s_{n+1}$  is.

You might argue that there still is a first pain state that is different from the baseline. Since the victim can tell that there is no difference between  $s_n$  and the baseline, but is uncertain whether  $s_{n+1}$  is different from the baseline, there is a perceptible difference between these two states.<sup>19</sup> However, the argument from indeterminacy applies once more. There might be second-order indeterminacy. There might be a range of pain states for which it is indeterminate whether it is indeterminate whether they are different from the baseline. Therefore, there might be no  $n$  such that there is no perceptible difference between  $s_n$  and the baseline, while it is indeterminate whether  $s_{n+1}$  is. These arguments could be endlessly repeated for higher and higher orders of indeterminacy. Importantly for our purposes here, any claim to the effect that there must be a first pain state for which it is  $k$ -order indeterminate whether it is perceptibly different from the baseline (where  $k$  is a natural number) may be countered with a claim about  $k+1$ -indeterminacy. For this reason, argument #3 does not establish that there necessarily is a first pain state that is perceptibly different from the baseline.

Carlson (1996) and Norcross (1997) gives an argument for why there still must be a first pain state that is perceptibly different from the baseline. This is the argument. Say that it is *superindeterminate* whether a pain state is perceptibly different from the baseline just in case there is *indeterminacy at some level* about whether the pain state is perceptibly different from the baseline. Then, there must be a pain state such that  $s_n$  is perceptibly different from the baseline, while it is superindeterminate that  $s_{n+1}$  is. There is no possibility that it is indeterminate whether it is superindeterminate that  $s_{n+1}$  is perceptibly different from the baseline. Cases of such indeterminacy are per definition cases of superindeterminacy. As Norcross (1997) puts it: “The postulation of indeterminacy about superindeterminacy is self-defeating” (144).

This argument is misleading. Contrary to appearances, indeterminacy about superindeterminacy is not self-defeating. To claim that there is superindeterminacy is simply to claim that there is indeterminacy at some level. And for any given level, it could be indeterminate whether there is indeterminacy at this level. For

---

where each triangulable difference involves barely noticeable pain, there must instead be plenty of triangulable differences.

<sup>19</sup> Compare the earlier discussion on three-valued logics.

illustration, consider a situation where the victim has three buttons in front of him, labelled “Yes”, “No” and “Indeterminate”. He is asked whether the current pain state is perceptibly different from the baseline, and after consideration he presses “Indeterminate”. Next, he is asked whether it is indeterminate whether the current pain state is perceptibly different from the baseline, and he once more presses “Indeterminate”. Successively, he is asked about higher and higher orders of indeterminacy, and for each order, he truthfully answers that it is indeterminate by pressing the appropriate button. This might seem to be a case of superindeterminacy, but it is not. There is superindeterminacy only if *it is true* that there is indeterminacy at some level. That is, there is superindeterminacy only if there is a pressing of “Indeterminate” followed by a pressing of “Yes” somewhere in the sequence. However, in the case under consideration, this never happens.

Whenever someone argues that there must be a first pain state for which is it superindeterminate whether this pain state is perceptibly different from the baseline, I would recommend the following argumentative strategy: First, clarify that the claim that there is superindeterminacy simply is the claim that there is indeterminacy at some level  $k$  about whether the pain state is perceptibly different from the baseline. Second, argue that it cannot be necessarily true that there is such a pain state since there might be  $k+1$ -level indeterminacy.<sup>20</sup> In general, you could point out that the mere possibility of never-ending indeterminacy entails that both argument #3 and the argument from superindeterminacy fail.

## Conclusion

Some important details of Kagan’s (2011) argument that it is a conceptual truth that there are no imperceptible difference cases are ambiguous. Nefsky (2012), considering what I have called argument #1, argues that Kagan mistakenly takes sorites arguments to be available in a *reductio* against the inductive premise. She is correct that argument #1 is not logically compelling – there are indeed alternatives to the rejection of its inductive premise. Still, our best theories of vagueness support Kagan on this point. The inductive premise is mistaken, and the flipping of some switch does make a difference as regards whether the victim is in pain or not. Moreover, even if it seems incredible to suggest that we should reject the inductive premise, the alternatives are even less attractive.

However, there is another problem with argument #1. Suppose we agree that it is a conceptual truth that there is some  $n$  such that the victim is in no pain in  $s_n$  while he

---

<sup>20</sup> Carlson (1996) is aware that the argument from superindeterminacy might be flawed. Right after presenting his argument, he writes “Perhaps the argument in the last paragraph is faulty”, and then goes on to suggest a rule for decision-making in cases of cyclical preferences which does not assume that triangulation always is possible.

in pain in  $s_{n+1}$ . This would not entail that the victim can perceive the difference between  $s_n$  and  $s_{n+1}$ . We cannot infer truths about perception from truths about concepts.

Argument #2 avoids this problem but runs into a different one. It assumes, questionably, that relations such as “cannot perceive the difference between” are transitive – questionably, because this assumption goes against the standard view and requires some careful defending.

There is a third argument to hand that, if successful, proves that there necessarily is a first pain state that is perceptibly different from some baseline state, and by extension that flipping a switch in HARMLESS TORTURERS has the equivalent of 1/1000 chance of triggering excruciating pain. Still, it is not successful. Since there might be never-ending indeterminacy in our perceptions, there is a possibility that there is no first pain state that is determinately different from the baseline.

Since argument #1, #2 and #3 fail, we still lack a conceptual argument showing that imperceptible difference cases are impossible. Ultimately, it may turn out that they are impossible. I have not produced arguments to the contrary. But until one is produced, we should consider imperceptible difference cases a possibility.

As a final note, I think the prospects of showing that there are sharp boundaries in our perceptions using conceptual arguments are bleak. However, even if it would turn out that all such arguments fail (and not just the ones I have considered), this does not necessarily entail that we must reject the expected utility approach. Instead, we might question THE PERCEPTIBILITY PRINCIPLE.



## 9. Denying the Description II

If what I said in the previous chapter is correct, we still lack a convincing argument showing that in every imperceptible difference case there is an act that triggers harm. (A) might still be true; that is, there could be cases where no act of  $\varphi$ -ing makes a perceptible difference to the harming of anyone, and yet, if enough people  $\varphi$ , someone will perceive a difference in harm. We cannot rule out that imperceptible difference cases exist, and by extension we cannot rule out that non-threshold cases exist.<sup>1</sup> This means that the expected utility approach still generates counterexamples. It gives counterintuitive verdicts about what reasons we have in non-threshold cases (as does any approach that similarly relies on SIMPLE). Luckily, there is another option for the expected utility theorists. They can try to demonstrate that THE PERCEPTIBILITY PRINCIPLE (B) is false. They can affirm that imperceptible effects can also be harms.

There is a convincing argument that shows this. I have in mind is Zach Barnett's (2018) *no free lunch argument*. If this argument bears scrutiny, all alleged non-threshold cases are cases where each act makes a morally relevant difference. If that is so, non-threshold cases pose no problem for the expected utility approach, or indeed for any approach that relies on SIMPLE. They are simply not possible.

The expected utility approach does, however, face another problem. It fails to properly explain our objective reasons in threshold cases. This point is particularly salient when it comes to considerations about blameworthiness and causation. But before we look into this issue, let us consider the no free lunch argument.

### No Free Lunch

Barnett (2018) argues that tiny differences can be morally relevant. His argument starts with the following case (which resembles DROPS OF WATER).

---

<sup>1</sup> I will go back to using the term “non-threshold cases” instead of “non-triggering cases”, which is the term Nefsky (2012) uses, and which I used in the previous chapter. These terms are interchangeable.

STAIRCASE: The 10,000 travellers are suffering from intensely painful thirst. They come upon a massive, 10,000-step staircase. Each step contains a partially filled canteen. The canteen on Step 1 contains 1 drop; the canteen on Step 2 contains 2 drops; and so on.

The travellers manage to arrange themselves on the staircase, with one traveller per step. Just before they take a drink, the traveller on Step 1 proposes an idea: “Wait! I was thinking... What if you all just moved down one step, and I moved up to the top?” She proceeds to explain that on this proposal, no one would be harmed (for all others forfeit only one drop), while she would benefit.

(Barnett 2018: 8-9)

Here, it seems that if the travellers agree to this suggestion, we have created a free lunch – or a free canteen of water, as it were. If the travellers move as suggested, the one sent to the top can enjoy a full pint of water while no one’s suffering is made worse. However, this conclusion seems implausible. As Barnett argues: “shuffling people around on the staircase does not seem likely to improve matters” (9), at least not if we assume that they have equal tolerance for thirst and that there are no other morally relevant differences between them. Barnett then concludes that “Even tiny contributions are morally significant” (9). If he is right, THE PERCEPTIBILITY PRINCIPLE is mistaken. It is possible for one drop of water to be morally relevant even if it does not make a perceptible difference.

To be precise, Barnett does not argue against THE PERCEPTIBILITY PRINCIPLE, but against the following principle:

NO SMALL IMPROVEMENT: The addition or subtraction of a single drop of water to/from someone’s canteen cannot (on its own) make her suffering better or worse.

(Barnett 2018: 5)

Still, his argument generalises to any non-threshold case. For instance, it could be used to show that it is false that no single flipping of a switch makes the victim’s suffering better or worse in HARMLESS TORTURERS.<sup>2</sup>

The strength and beauty of Barnett’s argument lies in its simplicity. It does not involve sorites-style reasoning, it does not assume that relations like “feels the same

---

<sup>2</sup> This is true whether you use the version of HARMLESS TORTURERS I presented in the introduction, where the victim is in mild pain at the start of the way, or Kagan’s version, where the victim is in no pain at the start of the day, used in the previous chapter.

as” are transitive, and it does not presuppose that triangulation works.<sup>3</sup> As Barnett makes clear, it relies simply on the following assumption:

[PARETO:] if one person’s suffering is relieved substantially while no one else’s suffering is affected, then the total suffering is reduced.

(Barnett 2018: 9)

Given this assumption, NO SMALL IMPROVEMENT straightforwardly entails the implausible verdict that “shuffling people around on the staircase” improves the situation in STAIRCASE. The 9,999 people who move down one step only lose one drop of water each, and according to NO SMALL IMPROVEMENT, one drop of water cannot make anyone’s suffering better or worse. So, we have a situation where one person’s suffering is relieved substantially while the suffering of each of the remaining 9,999 persons is not affected. *Per* PARETO, this means that the total suffering is reduced when the 9,999 people move down one step and the person on step one moves to the top. This is implausible, so either PARETO or NO SMALL IMPROVEMENT is mistaken.

You might be tempted to accept that the total suffering is reduced if the 9,999 people move down one step and the person at the bottom of the staircase moves to the top. After all, you might think, one person gets a full canteen of water extra, and the others only lose a single drop. Still, there are reasons to think that this verdict is mistaken. For one thing, suppose the 10,000 people arrange themselves randomly on the staircase when they first come upon it. If we think that we can reduce suffering by moving the person at the bottom to the top and everyone else down one step, it does not matter how they first happen to arrange themselves on the staircase. It will always be slightly better if the person at the bottom had been at the top and the others at one lower step. This verdict seems strange. Moreover, if the total suffering is reduced when the person at the bottom moves to the top and everyone else move down one step, we could repeat this manoeuvre, and thereby repeatedly reduce the resulting suffering when people eventually drink the water in their canteens. Theoretically, we could repeat this manoeuvre 10,000 times until everyone stands in their original position, having exactly the same amount of water as they did before we started shuffling people around, and claim that we have reduced suffering by doing so. This is patently untrue. The correct verdict must be that we do not reduce suffering by moving the person at the bottom to the top and

---

<sup>3</sup> Broome (2019) suggests another argument that aims to show that harms can be imperceptible, building on an argument suggested by Parfit (1984). Broome’s argument presupposes (i) that the betterness relation among pains is vague (and not incommensurate), (ii) that supervaluationism gives the correct account of vagueness, and (iii) that each relevant sharpening is a complete ordering of preferences. I prefer Barnett’s argument, primarily because it assumes much less than Broome’s.



everyone else down one step, and therefore that either NO SMALL IMPROVEMENT or PARETO is incorrect.

Those wedded to NO SMALL IMPROVEMENT will presumably argue that PARETO fails. Perhaps they will point out that PARETO presupposes that the group's suffering is unchanged unless some individual's suffering changes. PARETO, they might say, presupposes what we might call

[INDIVIDUALISM: if] the suffering is the same for each person, their total suffering [...] is also unchanged.

(Barnett 2018: 7)

Barnett introduces the principle I have labelled INDIVIDUALISM as an “important clarification” (7). However, as Erik Carlson, Magnus Jedenheim-Edling and Jens Johansson (2021) point out, INDIVIDUALISM is a substantial principle that requires careful defence.<sup>4</sup> Someone might argue that even if it is true that no one's suffering is increased when the travellers on step 2 to 10,000 move down one step, their total suffering increases. That is, the *group* consisting of these 9,999 people moving down a step suffers more than it would have done had its members not moved, and this is true even if no individual member of the group suffers more. Moreover, someone might continue, this increased group suffering explains why the total amount of suffering is unchanged when people are shuffled around on the staircase: the person moving to the top suffers less, and the group of people moving down one step offset this gain by suffering more.

The potential failure of INDIVIDUALISM might be a problem for those who want to show that NO SMALL DIFFERENCE is false. Still, from the point of view of the expected utility theorist, the potential failure of INDIVIDUALISM is not a big problem. Even if we reject INDIVIDUALISM, Barnett's argument will still show that the loss of one pint's worth of water makes a morally relevant difference even if this loss is distributed so that 9,999 people lose only one drop each. It makes a morally relevant difference for *the group* of people moving down the ladder even though it makes no morally relevant difference for any individual. If INDIVIDUALISM turns out to be unfounded, we can use this result to show that you have a reason to donate your pint in DROPS OF WATER. You have this reason since one pint of water makes a morally relevant difference for the group of suffering people.<sup>5</sup>

---

<sup>4</sup> I have borrowed the labels “pareto” and “individualism” from Carlson et al. (2021).

<sup>5</sup> Carlson et al. (2021) suggest a way to recast Barnett's argument in a way that does not presuppose INDIVIDUALISM. If successful, this argument would show that NO SMALL DIFFERENCE is false without assuming INDIVIDUALISM. However, their alternative argument presupposes that triangulation always is possible, which it is not (as I showed in the previous chapter).

If either conclusion is correct – that is, if it is true either that the addition or subtraction of a single drop of water to/from someone’s canteen can make her suffering better or worse, or that the addition or subtraction of a single drop of water to/from the canteen of each of the 9,999 people makes this group’s suffering better or worse even though the suffering of no particular person gets better or worse – we do not even need to appeal to expected utility to show that you have a reason to add your pint to the cart in DROPS OF WATER.<sup>6</sup> The situation here is not like that in FACTORY-FARMED CHICKEN, where each purchase *risks making* a difference to the outcome. Rather, each pouring *does make* a morally relevant difference to the outcome, and this is why you have a reason to pour your pint into the cart. It is a simple case where some harm will occur if you act in one way, but not occur if you act in another.

This point generalises to any non-threshold case. In any such case, each act of the relevant type makes a morally relevant difference, and therefore you have a reason to act in a certain way. For instance, flipping a switch in HARMLESS TORTURERS does make a morally relevant (albeit imperceptible) difference, and therefore each torturer has a reason not to flip his switch. Similarly, given that climate change is a non-threshold case, going for a single drive with a fossil fuel car makes a morally relevant difference, and therefore you have a reason not to go for such a drive. This means that non-threshold cases are not bona fide collective harm cases. In them, each act does make a morally relevant difference. In turn, this means that the inefficacy argument does not apply. In these cases, you have a reason to  $\varphi$  since doing so makes a difference to whether the collective outcome occurs.

## Problems in Threshold Cases

According to the standard view, the expected utility approach gives intuitively correct verdicts on the reasons we have in threshold cases, but runs into trouble in non-threshold cases. I think the standard view gets things the wrong way around. The expected utility approach gives the right verdict in so-called non-threshold cases, since in these cases each act makes a morally relevant difference. However, it fails to capture our intuitions accurately in threshold cases.

Others have contended that the expected utility approach runs into trouble in threshold cases. Mark Budolfson (2019) argues that the expected utility of doing

---

<sup>6</sup> In DROPS OF WATER, there are 10,000 people each of whom will receive a drop if you donate your pint. By contrast, in STAIRCASE, there are 9,999 people each of whom will lose a drop of water if they move around on the staircase. These differences should not bother us. We could easily amend STAIRCASE to show that one drop of water less makes someone suffer more; or alternatively, if you reject INDIVIDUALISM, that a group consisting of 10,000 people suffers more if each of its members gets one drop of water less.

your part in many real-life situations involving collective impact is much lower than expected utility theorists like Singer, Norcross and Kagan imagine. Take, for instance, a presidential election in which the electorate numbers 1,000 voters and there are only two candidates. Here, given a well-informed estimate that one candidate is likely to receive more votes than the other, the chances that your vote will make a difference to who wins the election are significantly lower than one in a thousand. Therefore, Budolfson argues, our reason for voting for the right candidate is much weaker than the expected utility approach indicates. Brian Hedden (2020) has also argued that we are entitled to distrust the expected utility approach because it gives the wrong verdict in cases involving infinities.

Budolfson's argument is unsound. As Hedden (2020) points out, it assumes that we have some knowledge of what will happen, and that therefore the expected utility theorist can adjust the likelihood of the outcome accordingly (and thereby the strength of the reason). There is no real problem for the expected utility theorist here. Yet, Hedden is right that we cannot trust the expected utility approach in cases involving infinities. I want to set that problem aside, however, and instead focus on a different one.

## **Expected Utility and Overdetermination**

The expected utility approach relies on an idea which, in some respects, reminds us of SIMPLE. It relies on the idea that what matters, in determining what reasons I have, is whether my act will make a difference to the occurrence of harm (or good). As in SIMPLE, this idea is notorious for generating counterintuitive results in cases of pre-emption and overdetermination (see e.g. Parfit 1984). Do these problems just disappear when we turn from objective to subjective reasons?

In one way, they do. Consider again FACTORY FARMED CHICKEN, and suppose that it turns out that 26 chickens were sold on that particular day, and that the butcher therefore ordered an additional batch of 25 freshly slaughtered chickens. In this case, my purchase made no difference to whether this order was placed or not. The butcher would have ordered the extra batch of chickens whether I bought a chicken or not. He orders an additional batch every time 25 chickens are sold. On a typical consequentialist model of the kind advocated by Singer, Norcross and Kagan, I had a subjective reason not to buy a chicken because, at the time, I did not know whether my purchase would make a difference. There was a risk that it would make a difference, bringing about harm, even though – as things turned out – it did not.

In another way, however, the problems do not disappear. Expected utility theorists will also say that if I had known all the facts, I would have realised that I had no future-suffering-of-chickens-related reason not to buy one chicken. Doing so would have made no difference to any future chicken's suffering. That is, they will say that I had no objective reason not to buy a chicken. I think this claim is false. When

expected utility theorists make it, they reveal their own “ethical anomie”, as Kutz (2000) puts it. Mistakenly, that is, they take one’s moral relations to the world to be essentially isolable. The correct view, I think, is that I had an objective reason not to buy one chicken. I had this reason because: there was a possibility that more chickens would suffer, and a possibility that no more chickens would suffer, and the outcome where more chickens suffer was such that it would be more secure were I to buy the chicken.

Not everyone would agree. But even if we follow Singer, Norcross, Kagan and others in holding on to the idea that what matters for what reasons I have is whether my act will make a difference to the occurrence of harm (or good), we run into other, related problems. For instance, given that we also hold the closely related idea that what matters for whether I cause an outcome is whether my act makes a difference to the occurrence of this outcome, we are obliged to say that no customer caused the additional batch of 25 chickens being hatched, raised and slaughtered under current factory farm conditions.<sup>7</sup> This seems a strange view to take. Surely, the suffering of these chickens occurred because of what the customers did. In addition, and perhaps more to the point,<sup>8</sup> we find that no customer is blameworthy for the suffering of these chickens. Objectively speaking, they had no reason to refrain from buying a chicken, because what they did made no difference to the suffering of any chicken. So, they cannot not be blameworthy for the future suffering of chickens. At most, they are merely blameworthy for performing an act that might have resulted in harm, but did not do so. This verdict seems strange.

This point might become clearer if we reconsider ASSASSINS. Recall that two assassins simultaneously and independently shoot one victim, and each shot is sufficient to kill the victim. In this case, the expected utility approach entails that each assassin had a subjective reason not to shoot, because doing so risked causing victim’s death, but also that, since, as things turned out, neither assassin caused the death of the victim, neither assassin was blameworthy for murder. At most, each was blameworthy for attempted murder. This seems incorrect. Each assassin caused the death of the victim, and each is blameworthy for his death. Likewise, each of the customers caused the additional batch of 25 chickens to be hatched, raised and

---

<sup>7</sup> We might think I only have an outcome-related reason to act in a certain way if it is the case that whether this outcome will occur depends on whether I act in this way, but still do not think that my act only was a cause of an outcome if it is the case that whether this outcome occurred depended on whether I performed this act. That is, we might think that there is more to causation than what is given by SIMPLE, but still think that what matters for what outcome-related reasons I have to  $\varphi$  is whether the occurrence of the outcome depends on whether I  $\varphi$  or not. Still, it is natural for consequentialists to think about causation in terms of SIMPLE. Whether you should  $\varphi$  is decided by the consequences  $\varphi$ -ing would have, where what these consequences are in turn is decided by something like SIMPLE.

<sup>8</sup> The expected utility approach is an account, not of causation, but of the consequences that are morally relevant.

slaughtered, and each customer is blameworthy for this, even in the case where a 26th chicken is sold.

The expected utility theorist might counter this objection by arguing that cases like FACTORY-FARMED CHICKEN do not involve overdetermination; they involve preemption. Lawford-Smith (2016) and Eriksson (2019) suggest this, in effect. The idea is that on a day on which 26 chickens are sold, it is the first 25 customers who cause the additional batch of freshly slaughtered chickens to be ordered. So, the first 25 customers had an objective reason not by a chicken, and are blameworthy for doing so. However, as long as we hold on to SIMPLE (or some version of it), this takes us from bad to worse. If SIMPLE is to give the right verdict in cases of preemption, we must either agree with Lewis (1973) that causation is a transitive relation, or focus attention on very fragile versions of the outcome. As I argued in Chapter 3, neither strategy successfully distinguishes contributions from counteractions. To see why appealing to the transitivity of causation does not work, consider the following situation: I try to persuade you not to buy a chicken, but you refuse to listen to my arguments and buy one anyway. If we agree with Lewis that causation always is transitive, we have to conclude that I caused you to buy a chicken by trying to persuade you not to do so. There is stepwise counterfactual dependence here. Had I not tried to persuade you, you would not have refused to listen to me. And had you not refused to listen to me, you would not have bought a chicken. So, my attempt to persuade you caused you to buy the chicken, and by extension I might have caused the future suffering of chickens. So, if we pick up on Lawford-Smith's and Eriksson's suggestion, we must conclude that I had an objective future-suffering-of-chickens-related reason not to try to persuade you to refrain from buying a chicken, and that I am blameworthy for the future suffering of chickens since I did try to persuade you. (To see why the fragility strategy does not work, see discussion in Chapter 3.)

## Conclusion

The inefficacy argument can be resisted either by rejecting the implication or by rejecting the description. The last three chapters have considered the prospects of the second of these options. Can we show that there is an act that makes a difference in collective impact cases? Most of the discussion in the literature revolves around non-threshold cases like DROPS OF WATER and HARMLESS TORTURERS. In these, no act makes a perceptible difference to the suffering of anyone. Kagan (2011), Norcross (1997) and others argue that some act makes a perceptible difference in all non-threshold cases. I have tried to show that their arguments for this view are mistaken. Others, like Barnett (2018), have argued instead that imperceptible harms *are* harms. I agree. If Barnett and I are correct, genuine non-threshold cases do not exist. In them, each act makes a morally relevant difference.

The inefficacy argument still looks powerful when applied to threshold cases like voting, ASSASSINS, FACTORY-FARMED CHICKEN and THE LAKE. Expected utility theorists accept the conclusion that you have no objective outcome-related reason to act in the relevant way in these cases. They have recourse to the claim that you have a subjective reason. Given your limited knowledge, you could not know in advance that your act would make no difference to the outcome, and given this you did have a reason to act in the relevant way, since it might have turned out that your act made a difference to the outcome.

I disagree. You do have an objective outcome-related reason in cases like voting, ASSASSINS, and so on. That you have such a reason is most obvious if we put reasons to one side for a moment and consider blameworthiness. If no assassin had an objective reason not to shoot the victim, no assassin was blameworthy for the victim's death. The most that each assassin can be blamed for is attempted murder. This seems wrong. The victim died, and the assassins are blameworthy for murder, not merely for attempted murder.

This shows that we need, not just an improved account of objective outcome-related reasons (like REASON), but also a more compelling account of the conditions under which you are blameworthy for an outcome. I will seek to provide an account of the latter in Part Two.



## Part Two

# Blameworthiness and Causation

When are you blameworthy for an act, omission or outcome? Could you be blameworthy for the men's continued suffering if you refuse to donate your pint in drops of water? Or could you be blameworthy for the collapse of the ecosystem in the lake if you use the hazardous paint? You could. Roughly, you are blameworthy for X – where X is an act, omission or outcome – if and only if a poor quality of will of yours in relation to X was a cause of X, or so Touborg and I will suggest. In this context, to have a poor quality of will towards something is, roughly, to care too little about it, and being blameworthy for something means that others are warranted in reacting negatively to you because of this thing. For instance, in drops of water, others are warranted in being angry with you because of the men's continued suffering if and only if your caring too little about the men's suffering (and its alleviation) is a cause of their continued suffering. If your inadequately caring attitude made you keep your pint, this would be the case. This suggestion is presented in Chapter 11, "You Just Didn't Care Enough", and further elaborated in Chapter 12, "Elaborating BLAMEWORTHINESS FOR". The resulting account is applied to further standard cases in Chapter 13 "Applying BLAMEWORTHINESS FOR". Finally, in Chapter 14, "Moral Entails Causal", Sartorio's arguments for thinking that you might be blameworthy for an outcome even though you did not cause it are examined and rejected.





## 10. Blameworthiness For

Before presenting my proposal (jointly developed with Touborg) concerning when an agent is blameworthy for an act, omission or outcome, I would like to say a few words about what it means to be blameworthy *for* something.

Let me start by saying what it does not mean. On one understanding, to be blameworthy is to be – partly or wholly – a bad person. A bit more carefully, on this view, you are blameworthy if and only if you have a bad quality of will, where having a bad quality of will amounts to something like having bad motives, intentions or dispositions, or simply being a bad person or having a bad character. Being blameworthy, in turn, typically means that you are the fitting object of what P. F. Strawson (2008/1962) calls “negative reactive attitudes”, like resentment, indignation, or (in the case of self-blame) guilt. Proponents of this view include Angela Smith (2005, 2008), T. M. Scanlon (2008),<sup>1</sup> Pamela Hieronymi (2008) and Matthew Talbert (2012, 2019). Neil Levy (2005) and Talbert (2012, 2019) call the view *attributionism*, and I will follow suit.<sup>2</sup>

According to attributionists, it does not matter what you are blameworthy for.<sup>3</sup> Your actions and omissions, and their outcomes, might serve as outward evidence of your quality of will, but they are ultimately irrelevant to whether you are blameworthy. Consider, for instance, Michael Zimmerman’s (2002) case in which each of two assassins tries to kill someone but one fails due to factors beyond his control.

[PASSING BIRD:] Suppose that George shot at Henry and killed him. Suppose that Georg shot at Henrik in circumstances which were, to the extent possible, exactly like those of George (by which I mean to include what went on “inside” the protagonists' heads as well as what happened in the “outside” world), except for the fact that Georg's bullet was intercepted by a passing bird (a rather large and solid bird) and Henrik escaped injury.

(Zimmerman 2002: 560)

---

<sup>1</sup> Scanlon uses a wider notion of blame than Strawson, but is nevertheless usually referred to as an attributionist.

<sup>2</sup> For an overview of the varieties of attributionism, see e.g. Talbert (forthcoming).

<sup>3</sup> Again, the exception here is Scanlon.

Attributionists would say that George and Georg are equally blameworthy. It does not matter that George is guilty of murder while Georg only is guilty of attempted murder. The fact that Henry was killed while Henrik survived might direct our attention to George's bad quality of will rather than to Georg's, with the likely result that we will tend to blame George more than Georg. However, if we consider the situation carefully, we will realise that both George and Georg are equally blameworthy. They are equally bad persons.

A related view is that you are blameworthy if and only if you have a bad quality of will *and* satisfy a normative competence condition. Susan Wolf (1990), for instance, argues that you are a fitting object of reactive attitudes only if you "know the difference between right and wrong" (382).<sup>4</sup> She considers the case of Jojo, the son of an evil sadistic dictator. Jojo grows up shadowing his father, who routinely imprisons, tortures and murders citizens on a whim. As a result, he becomes an evil sadistic dictator himself, much like his father. Attributionism entails that the adult Jojo is a fitting object of our reactive attitudes. As an adult, he has bad motives, intentions, dispositions, and so on. However, Wolf argues, the adult Jojo is not an appropriate object of such attitudes. His upbringing made him incapable of telling right from wrong, and as a result he is exempted from blame. To take a different example, on Wolf's view, racists who grow up in a racist society might not be blameworthy for being racists. They simply never learned right from wrong. Attributionists would disagree. Talbert (2012) argues, for instance, that a victim of racialised crime is warranted in blaming his aggressors, and that this is true even if these aggressors grew up in a racist society.

Even if you agree with Wolf that it is only agents who can tell right from wrong who can be blameworthy, you might also agree with attributionists that questions about what you are blameworthy for are irrelevant. Thus you might agree that George and Georg are equally blameworthy. There is no difference between George and Georg in their normative competence.

The result that consequences do not matter for blameworthiness can certainly seem strange. For one thing, we are used to a legal system where there is an important difference between murder and attempted murder. Moreover, discussions of who is blameworthy, and for what, are pervasive in everyday life as well as in academia. As Joel Feinberg writes:

---

<sup>4</sup> To be precise, Wolf says that you are the kind of agent that can be morally responsible if you (a) can govern your actions with your desires, (b) you can govern your desires with your deep self, and (c) your deep self is *sane* (that is, deep down, you can tell right from wrong). Roughly, we get attributionism if we remove (c). If you satisfy (a) and (b), your actions tell us something about your quality of will. Wallace (1994) would agree with Wolf that there is some kind of normative competence requirement on being blameworthy. He suggests that you must have "reflective self-control", which is "the power to grasp and apply moral reasons and ... to control or regulate [their] behavior in the light of such reasons" (157).

“[H]is fault” judgements, as I will call them, are important and ubiquitous in ordinary life. Historians employ them to assign blame for wars and depressions; politicians, sportswriters, and litigants use them to assign blame for losses.

(Feinberg 1970: 187)

At this point, I would like to introduce a distinction between *blameworthiness* and *blameworthiness for*. Attributionism and Wolf’s view are instances of the former, while Feinberg seems to have the latter in mind.<sup>5</sup> Assessments of *blameworthiness* are evaluations of the character, the moral fibre, or quality of will of a person (perhaps in combination with some further condition), whereas assessments of who is *blameworthy for* something are assessments of whose fault this something is, if anyone’s. In the former, causation and consequences play no role.<sup>6</sup> In the latter, causation and consequences are essential.

The following might help to elucidate this distinction. Imagine a situation where someone accuses you of being blameworthy, and does so without further explanation. How would you react? Perhaps you would ask them what they are blaming you for. If they reply that they are not blaming you for anything, you will presumably be perplexed. In fact, you might even become angry with them for blaming you for nothing in particular. If this is how you react, you are probably thinking in terms of *blameworthiness for*. On this view, you cannot be blameworthy without being blameworthy *for* something. There must be some object of blame. Alternatively, you might react by denying that you are a horrible person, and ask what grounds they have for thinking that you are one. If this is how you react, you are probably thinking in terms of *blameworthiness*. You would probably agree with Smith (2008) and others that there is no important distinction to be made between bad agents and blameworthy agents.

This might also help to elucidate the distinction. While questions about *blameworthiness* can be answered by considering what dispositions, intentions or quality of will an agent has at one single point in time (or during one single time interval), answering questions about what you are blameworthy for requires that we consider two separate points in time (or two separate time intervals): one at which

---

<sup>5</sup> Accounts of *blameworthiness* also include, I think, so-called *mesh* or *real self* theories. These claim that you are blameworthy only if, roughly, you endorse the desires on which you act. According to Frankfurt (1971), for instance, you are only blameworthy for an action of yours if your first-order volitions (your action-guiding desires) are in conformity with a second-order volition (i.e. a desire that a certain desire of yours is action-guiding). Although Frankfurt talks about blameworthiness for actions, his main idea seems to be that these actions tell us something about your real self only when your real self is in control over your behaviour. In this way, what really matters on this view is the evaluative judgements of your real self. Watson (2004/1975) proposes a similar view.

<sup>6</sup> Perhaps better: causation and consequences play no more than the derivative role of serving as outward evidence of someone’s quality of will.

the action, omission or outcome occurs, and another at which the agent was in control over, or had a poor a quality of will toward (or something of the sort) that action, omission or outcome. Thus, when we are asked whether George is blameworthy, it is enough to consider his character, dispositions, intentions, and so on, at a certain point in time, or during a certain time interval. However, when we are asked whether George is blameworthy for Henry's death, we must consider two separate times: the time at which Henry died, and the time at which George was in control over whether Henry would live or die (or at which George had a poor quality of will towards Henry, or at which something of the sort transpired).

Note also that you could be blameworthy *for* being a bad person, or *for* having bad dispositions. This would be an instance of being blameworthy for an outcome, not a case of being blameworthy (*simpliciter*). To decide whether you are blameworthy for being a bad person or having bad dispositions, it matters how it came to be that you are such a person or have such dispositions. To decide whether George is blameworthy for having the disposition to kill Henry, for instance, it matters whether he was in control over his dispositions, whether he intentionally formed these dispositions, whether he formed them (or let them be formed) out of disregard for Henry, and so on. That is, to decide whether George is blameworthy for having the disposition to kill Henry, we have to consider two separate time slices or intervals: one at which he has this disposition, and another earlier time at which was in control over this disposition.

The account introduced in the next chapter is intended to explain when you are blameworthy for something. It is not about whether you are a bad person or have a bad character. It sets out the circumstances, or conditions, under which you are blameworthy for an act, omission or outcome. This is not to say that questions about blameworthiness *simpliciter* are irrelevant or uninteresting. They are just not what (Touborg and) I are talking about.

There are two main approaches to the question of *blameworthiness for* in the literature. The first goes back to Aristotle. It claims an agent is blameworthy for an action, omission or outcome just in case she satisfies a control condition and an epistemic condition. On views of this sort, you are only blameworthy for what you did, or brought about, if you were in control of what you did, or brought about, and if you knew its moral import (or should have done so). John Martin Fischer and Mark Ravizza (1998), David Brink and Dana Nelkin (2013) and Carolina Sartorio (2016) propose views along these lines.<sup>7</sup>

On the second approach, essentially, you are blameworthy just in case you intentionally act in a way that can be criticised, or intentionally cause a bad outcome.

---

<sup>7</sup> Fischer & Ravizza and Sartorio focus on the control (or “freedom”) condition of moral responsibility, largely setting the epistemic condition aside. For a collection of papers that discuss the epistemic condition, see Wieland and Robichaud (eds, 2017).

A variant of this idea has it that you are blameworthy just in case you wrong someone out of disregard for them, or just in case a bad outcome occurs because you do not care enough about the consequences of your actions. Matthew Braham and Martin Van Hees (2012) and Gunnar Björnsson (2017a, 2021) advocate such views. The account we propose belongs to this second category.

Before I introduce our account, there is an important issue to address. Some have argued that questions about what you are blameworthy for are irrelevant (e.g. Zimmerman 2002; Enoch & Marmor 2007; Peels 2015). In particular, they have argued that what you are blameworthy for makes no difference for how blameworthy you are. In what remains of this chapter, I argue that this view is mistaken. My main argument is as follows. If we understand blameworthiness in terms of warranted, or fitting, reactive attitudes, George is more blameworthy than Georg. Henry's parents, for instance, are warranted in reacting more negatively to George for killing Henry than Henrik's parents are in reacting to Georg for attempting to kill Henrik. This means that while George and Georg are equally bad persons, it is fitting for at least some people to blame George more than Georg, which amounts to saying that George is more blameworthy than Georg. Hence, it matters whether you are blameworthy for killing someone, or merely for trying to do so. This point generalises to other cases. Thus if a drunk driver accidentally runs over and kills a child, we are warranted in blaming this driver more than a driver who "merely" drives drunk but luckily does not run over anyone. We could make similar points using other understandings of blameworthiness, such as Scanlon's.

Besides presenting Zimmerman's argument and my misgivings about it, I will discuss some objections to the idea that the fact people are warranted in blaming George more than Georg shows that George is more blameworthy than Georg. But first things first. Here is Zimmerman's argument for thinking that questions about what you are blameworthy for are irrelevant.

## Does Scope Count for Nothing?

Zimmerman (2002) argues that "judgements about responsibility *for* something are essentially otiose" (570). Commenting on George and Georg in *PASSING BIRD*, he writes:

They are equally responsible; if George is deserving of a particular reaction, then Georg is deserving of the very same reaction. This indicates that whether there is something *for* which one is responsible is immaterial; all that matters, fundamentally, is whether one *is* responsible. Degree of responsibility counts for everything, scope for nothing, when it comes to such moral evaluation of agents.

(Zimmerman 2002: 569)

While Georg is morally responsible for fewer things than George, they are both equally responsible, or so Zimmerman argues. Here, to be morally responsible for an outcome is to be (in the negative case) blameworthy or (in the positive case) praiseworthy for this outcome.

I disagree that judgements about responsibility *for* something are essentially otiose. Even if questions about the scope of responsibility do not determine, or help to determine, someone's degree of blameworthiness, it is still important to have an accurate account of who is blameworthy for what – for instance when we are settling legal disputes, when we are trying to sort out who is to blame for wars and economic depressions, and in general when we try to understand everyday discussions about who is to blame for what. At most, Zimmerman's argument establishes that questions about what you are blameworthy for are irrelevant to deciding your degree of blameworthiness. It does not establish that such questions are otiose. However, I would go further. I also disagree that the scope of responsibility does not matter for the degree of responsibility. If we have a firm grip on what blame and blaming is, we will see that we are warranted in blaming George more severely, and for a longer period, than we are in blaming Georg. Or, so I will argue.

Zimmerman's argument for thinking that scope counts for nothing is part of his widely discussed argument against moral luck.<sup>8</sup> The argument goes roughly like this. You are only blameworthy for things in your control. Georg was not in control of the fact that his bullet was intercepted by a bird, so this fact cannot affect his blameworthiness. Therefore, he is as blameworthy as George is.

This might seem counterintuitive. You might think Georg is not particularly blameworthy, because he did not kill anyone. If that were right, Zimmerman's claim that George is as blameworthy as Georg would entail that George is not particularly blameworthy. However, George does seem blameworthy. After all, he voluntarily killed Henry. In order to avoid this counterintuitive verdict, Zimmerman argues that Georg is blameworthy roughly to the degree that murderers typically are. But to establish this, he has to show that there is something in virtue of which Georg is blameworthy.

Zimmerman first suggests that Georg is blameworthy because he tried to kill Henrik. However, he soon rejects this suggestion. Georg was not in control of whether he tried to kill Henrik, so he cannot be blameworthy for doing that. Why was he not in control of whether he tried to kill Henrik? Well, Zimmerman says, in some counterfactual scenario, some truck might have pulled up in front of Henrik blocking Georg's firing line, or Georg might have sneezed immediately before pulling the trigger and therefore missed his only opportunity to try to shoot Henrik.

---

<sup>8</sup> The discussion of moral luck goes back at least to Nagel (1979) and Williams (1981). Zimmerman's argument can also be found in his (2011) *The Immorality of Punishment*.

This means whether Georg tried to kill Henrik was not in his control. It was a matter of luck.

Instead, Zimmerman considers whether Georg is blameworthy for “his being such that he would have freely killed Henrik, given the opportunity” (2002: 564). That is, he considers whether Georg is blameworthy in virtue of having the disposition to kill Henrik. Zimmerman rejects also this suggestion. Georg lacked control over the dispositions he had, so he cannot be blameworthy for having these dispositions. (Zimmerman does not explain why he lacked control over the dispositions, but perhaps nature and nurture play important roles here).

Rather, Zimmerman finally suggests, Georg is blameworthy in virtue of being such that “he would have freely killed Henrik because he would have freely chosen to shoot him, had he had the cooperation of certain features of the case” (Zimmerman 2002: 564). That is, Georg is blameworthy because there is some counterfactual world in which he has the disposition to kill Henrik, and as a result kills Henrik. (One might ask why the blameworthiness here is not undermined by the fact that Georg was not in control of whether there was some counterfactual world in which he voluntarily killed Henrik. I will come back to this issue.)

As Zimmerman notes, this “opens up the floodgates [...] when it comes to ascriptions of responsibility” (2002: 570). To the extent that counterfactual Georg is blameworthy, actual Georg is too. The only thing preventing Georg from being blameworthy for everything that anybody has ever been blameworthy for, and everything that anybody ever could have been blameworthy for, is that Georg must still be Georg in these counterfactual worlds. Georg is not, for instance, blameworthy for what Hitler did. In no counterfactual world is Georg Hitler. The same, incidentally, goes for George.

The same reasoning can be applied to praiseworthiness. Georg and George are not only blameworthy in virtue of what they do in counterfactual worlds. They are praiseworthy in virtue of what they do in counterfactual worlds. So, there are quite a few things on their moral ledgers.

### **Why Zimmerman’s Argument Fails**

Nathan Hanna (2014) and Robert Hartman (2017) argue that counterfactual views of blameworthiness like Zimmerman’s have implausible implications. As Hartman points out, they entail that “agents may be praiseworthy or blameworthy in virtue of events that are radically different from the kind for which they are praiseworthy and



blameworthy in the actual world” (2017: 65).<sup>9</sup> For illustration, he considers the following case:

Suppose that actual Ben is an average person, neither very good nor bad. Suppose also that although his parents do not get on a particular plane in the actual world, they get on it in the majority of nearby possible worlds. And in those close possible worlds, his parents die in an explosion on the plane due to modally resilient technical errors by the maintenance crew. In these close worlds, Counterfactual Ben forms the malicious desire to harm the people who are responsible and kills several people who work for the airline.

(Hartman 2017: 66)

Intuitively, Actual Ben is less blameworthy than Counterfactual Ben is. Counterfactual free actions and dispositions that differ radically in moral quality from your actual actions and dispositions do not make you blameworthy. At least, they do not make you blameworthy to the same high degree. Zimmerman’s view entails that Actual Ben is as blameworthy as Counterfactual Ben, however. Actual Ben is blameworthy because there are nearby possible worlds where he would freely commit murder. Surely, this is counterintuitive, Hartman says.<sup>10</sup> Hanna (2014) makes the same point, albeit in relation to another example.

Talbert (forthcoming) agrees that counterfactual views like Zimmerman’s have counterintuitive implications. He suggests, however, that we can avoid these if we hold on to the idea that your degree of blameworthiness is decided by your *actual* dispositions. That is, he argues that we should not accept the final step in Zimmerman’s argument. Refusing to take this step, we end up with attributionism. You are blameworthy if your actual dispositions (character, motives, quality of will) are morally deficient. This view gives the intuitively correct verdict that Actual Ben is not blameworthy while Counterfactual Ben is. Actual Ben’s dispositions are good enough, while Counterfactual Ben’s are not. But although Talbert rejects Zimmerman’s counterfactual view, he agrees with the idea that judgements about what you are blameworthy for are essentially otiose.

I think Zimmerman’s argument is unpersuasive, and that both Zimmerman and Talbert are wrong to think that judgements about what you are blameworthy for are essentially otiose. As a first pass, I would like to point out that if you agree with

---

<sup>9</sup> Besides Zimmerman, Hartman also directs his argument against Enoch and Marmor (2007) and Peels (2015), who present similar argument as Zimmerman does. Peels spells out the counterfactual view very clearly.

<sup>10</sup> Hartman spells out the example in terms of what happens “in the majority of nearby possible worlds” since Peels (2015) tries to avoid a similar objection by appealing to what you do in the majority of nearby possible worlds. If you are disposed to perform terrible acts in the majority of nearby possible worlds, Peels argues, you are blameworthy also in the actual world.

Zimmerman that you are only morally responsible for things in your control, you really should not be very satisfied with the view he eventually advocates. While he starts with the idea that you are only morally responsible for things in your control, he ends up with a view according to which control is irrelevant for moral responsibility. For instance, Actual Ben is responsible in virtue of the dispositions that Counterfactual Ben has (which is killing people out of a lust for revenge), and this is true on Zimmerman's counterfactual view even though Actual Ben has no control over the dispositions of Counterfactual Ben. This does not show that Zimmerman's view is incorrect. It does show that the motivation for this view is unfounded. If we take seriously the idea that things outside your control do not influence your degree of blameworthiness, we should conclude that the fact that you would have freely performed some good or bad action under counterfactual circumstances does not influence your degree of blameworthiness. What you would freely have done under counterfactual circumstances is beyond your control.

To assess Zimmerman's counterfactual view and the attributionist view, we must first obtain a firm grasp on what, exactly, blameworthiness is, and more generally on what moral responsibility is. On a standard view, to be morally responsible is to be the appropriate object of what Strawson (2008/1962) calls "reactive attitudes", such as resentment, indignation, forgiveness or gratitude. On this view, to blame someone is simply to direct such attitudes towards this person. Something like this view is accepted by most attributionists.<sup>11</sup>

With this view in mind, it seems that George is more blameworthy than Georg. Consider Henry's and Henrik's mothers, for example, and assume that they have discovered what George and Georg did. It would not be inappropriate for Henry's mother never to forgive George, nor would it be inappropriate if she resented him for what he did for the rest of her life. However, it would seem excessive of *Henrik's* mother never to forgive Georg or resent him for the rest of her life.<sup>12</sup> After all, Henrik is still alive. This indicates that there is a difference in the reactive attitudes it is appropriate to have towards George and Georg. George merits the more severe attitude, which means that he is more blameworthy.

The same conclusion seems to follow if, like Scanlon (2008), we take blame to be an adjustment of your attitudes to someone made in response to an impaired relationship with this person, and thus take someone to be blameworthy when it is fitting to adjust your attitudes to them in response to such an impairment. Since George killed Henry, his relationships with others (particularly Henry's parents and

---

<sup>11</sup> An exception is Scanlon, who sees reactive attitudes as typical examples of blaming reactions but argues that there are other kinds of blaming reaction.

<sup>12</sup> I assume here that she will live for quite a few years after Georg's attempt to kill Henrik.

friends) are impaired in a way that Georg's are not. George's actions played a significant role in some people's lives, whereas, luckily, Georg's actions did not.<sup>13</sup>

I think these points are in line with our moral practices: we take murderers to be more blameworthy than those who "merely" attempt to murder someone, and we take people who drink and drive, and as a result run over someone, to be more blameworthy than people who "merely" drink and drive. And so on.

Zimmerman (2002) does not subscribe to any of these views of blame and blameworthiness. Instead, he endorses a ledger view on which "to be morally responsible is to be such that there is an 'entry' in one's 'moral ledger' in light of some fact about oneself" (555). It seems to me that we need a firmer grip on what blame and blameworthiness consist in than Zimmerman provides in order to know what we should be looking for when considering the relevant examples. We need to know, for instance, whether we should concentrate on the reactions others are entitled to have, or instead seek to evaluate the character of the wrongdoers – or do something else entirely.<sup>14</sup>

Relatedly, it seems to me that Zimmerman's argument rests to some extent on an equivocation. Remember that he says that George and Georg "are equally responsible; if George is deserving of a particular reaction, then Georg is deserving of the very same reaction", and then concludes that "Degree of responsibility counts for everything, scope for nothing, when it comes to such moral evaluation of agents" (2002: 568). There is, however, an important difference between evaluating agents morally and asking what reactions to their conduct are merited. Zimmerman is right that when we evaluate George and Georg as agents, we conclude that they are equally bad. They have the same heinous dispositions, motives, intentions, characters, etc. However, he is wrong to say that George and Georg deserve blame to the same extent. They do deserve the same reactions *in virtue of the kind of agents they are*, but some of us are nonetheless warranted in reacting more negatively to George than we do to Georg *in virtue of what they did or voluntarily caused*.

---

<sup>13</sup> Scanlon (2008) discusses a case featuring two reckless drivers, one who happens to run over a child, and one who does not. He makes more or less the same point about that case as I do about PASSING BIRD. However, in making his point he writes in a somewhat ambiguous way. Talbert (2019) argues that Scanlon could accept a version of attributionism in which there is no such thing as consequential luck without contradiction. Maybe that is true. Even so, it seems that if you agree with Scanlon that blame is "a revised understanding of our relations with a person, given what he or she has done" (150), it seems quite natural to also think that different blaming responses are appropriate in the cases of George and Georg.

<sup>14</sup> Enoch and Marmor (2007), who present an argument that is similar to Zimmerman's, say little about what they take blame and blameworthiness to be. They reject the idea that you are morally responsible whenever you are the appropriate object of reactive attitudes, and also reject attributionism (or the "character-based" view, as they call it). They seem to rely on an intuitive grasp of what blameworthiness is. This is unsatisfactory. We need to know what we are talking about.

## Objections and Answers

I have argued that George is more blameworthy than Georg because we are warranted in blaming George more strongly than Georg. Some might object that although we are warranted in blaming George more strongly, it does not follow that he is more blameworthy than Georg. I will consider three such objections.

First, David Enoch and Andrei Marmor (2007) and Rik Peels (2015) hold that the degree to which we are warranted in blaming someone depends greatly on the evidence we have for thinking that this person has a poor quality of will. They add that when a bad outcome occurs as a result of someone's poor quality of will, this often provides us with evidence of the required kind. So, when some bad outcome occurs as a result of someone's poor quality of will, we are often warranted in blaming this person more than we would if the outcome had not occurred. I do not deny this, but I doubt that knowledge plays an important role in the case of George and Georg. In the case, what George and Georg did is openly acknowledged, so Henrik's parents and friends have the same evidence of Georg's quality of will as Henry's parents and friends have of George's quality of will, and therefore evidential differences should not matter in this case.<sup>15</sup>

Second, Enoch and Marmor (2007) and Peels (2015) also argue that we might be warranted in *overtly* blaming George more than Georg even if they are equally blameworthy. Overt blaming behaviour might be justified for all sorts of reasons, including educational reasons or deterrence. You might be justified in blaming young children even though they are not particularly blameworthy, if this is an effective way to teach them something important. Equally, it may be justifiable to punish wrongdoing harder than attempted wrongdoing because we want to educate the wrongdoers, and deter others from doing the same thing. It seems true that we sometimes are justified in blaming someone for educational or deterring reasons.<sup>16</sup> However, in the examples we are considering, forward-looking reasons do not appear to play an important role. For instance, in *PASSING BIRD*, we thought Henry's mother was warranted in reacting more negatively to George than Henrik's mother was in reacting negatively to Georg. Still, Henry's mother does not seem to have a stronger reason to encourage George to develop better dispositions and motives than Henrik's mother has to encourage Georg to do the same. If anything, Henrik's mother has stronger reasons to make Georg change his dispositions and motives. If

---

<sup>15</sup> Enoch and Marmor (2007) acknowledge that the argument from knowledge can be countered by appealing to other evidence in the case of consequential luck. They argue that the argument from knowledge works better in connection with circumstantial luck. There is more to say here, but doing so would take us too far afield.

<sup>16</sup> It seems to me that, if we go by forward-looking considerations alone, educating those who have tried to commit a crime but failed is as important as educating those who were successful in their criminal endeavour. It also seems to me that punishing attempted wrongdoing could have the same deterrent effect as punishing successful wrongdoing.

he retains them, the chances are that he will try to kill Henrik, who is still alive, at some later time.

Third, Peels (2015) argues that we often irrationally blame someone for something when, on further reflection, we would not blame that person – or, at least, not blame that person to the same degree. Thus, we might easily slip from (correctly) thinking that George brought about a more regrettable outcome than Georg did to (incorrectly) thinking that George is more blameworthy than Georg.<sup>17</sup> Peels states this objection briefly, but it seems pressing if we elaborate it along the following lines. Perhaps Henry's parents and friends are warranted in having stronger affective reactions than Henrik's parents and friends are. However, when we separate the reactive attitudes Henry's parents are warranted in having towards George (anger, resentment, and so on) from other affective reactions they are warranted in having (feeling grief and torment over the fact that Henry is dead), we see that they are not warranted in reacting more negatively to George than Henrik's parents and friends are in reacting negatively to Georg. That is, Henry's parents and friends are warranted in resenting George to the same degree as Henrik's parents and friends are in resenting Georg. If we mistakenly conflate the different kinds of affective reactions Henry's parents and friends are warranted in having, we might mistakenly think that they are warranted in reacting more negatively to George than Henrik's parents and friends are in reacting to Georg. This, one might argue, is what explains the intuition that George is more blameworthy than Georg (if we have that intuition).<sup>18</sup>

These are complicated matters. Undeniably, people do sometimes mix up what affective reactions they are warranted in having. Thus, Henry's parents and friends might project their grief and torment over Henry's death onto George when first learning about Henry's death. It is understandable if they do. However, this does not completely explain away the notion that George is blameworthy for what he did to a higher degree than Georg is for what he did. When things calm down and Henry's parents and friends get some perspective on things (this might take a while), they might realise that their reactions towards George was somewhat over the top, and instead place their grief and torment where it belongs. Even so, they might still think that they are warranted in resenting George for what he did to a higher degree, and for a longer time, than Henrik's parents and friends are in resenting Georg. After all, George killed Henry while Georg merely tried to kill Henrik. Similarly, from our third person perspective, we might at first conflate the affective reactions we (and others) are warranted in having towards George for what he did, and the affective reactions we are warranted in having towards the fact the Henry is dead. If we do, we might think that we are warranted in reacting even more negatively to

---

<sup>17</sup> Talbert (2017) puts a similar argument, but in relation to unwitting wrongdoing.

<sup>18</sup> Thanks to Björn Petersson for suggesting this elaboration of Peels' argument.

George than we are, which in effect means that we might think that George is even more blameworthy for what he did than he is. In that case, we might later realise that we are mistaken. We might realise that even though we are warranted in reacting negatively to George for what he did, we are not warranted in reacting as negatively as we first thought we were. Still, this does not show that George and Georg are equally blameworthy for what they did. After thinking things through, it still seems that we are warranted in reacting more negatively to George for killing Henry than we are in reacting to Georg for trying to kill Henrik.

One last point. We might slip from considering what someone is blameworthy for to considering whether someone is blameworthy *simpliciter*. We might for instance slip from considering whether – and to which degree – George is blameworthy for killing Henry to considering whether George is a bad person (that is, a person with a bad quality of will). And, we might do the same when it comes to Georg. If we only concentrate on the question of how George and Georg are as persons, we will find that both are equally blameworthy: they are equally bad persons. Still, if we do, we overlook the fact that Henry's parents and friends (and we) are warranted in resenting George more, and for a longer time, than Henrik's parents and friends (and we) are in resenting Georg. That is, in that case, we overlook that fact that George is more blameworthy for what he did than Georg is for what he did.

## Conclusion

In sum, to the extent that the consequences of our actions affect the reactive attitude that others are warranted in having towards us, what we are blameworthy for affects our degree of blameworthiness. There is more to say about this issue than I have said here. If it turns out that I am wrong, and that there is no such thing as consequential moral luck, this does not show that questions of what we are blameworthy for are uninteresting or otiose. It shows only that they do not matter for the degree of blameworthiness. They might still be of interest in settling legal or historical matters, or in accounting for everyday ascriptions of who is blameworthy for what.

I should also stress that I have not been arguing that judgements about *blameworthiness* are otiose. I have been arguing that the consequences of our actions can be relevant when we are seeking to determine which reactive attitudes others are warranted in having towards us. This does not exclude the possibility that bad dispositions, character, and so on, also might warrant certain reactive attitudes. In fact, although I have not argued for this point, I think this is quite plausible.



# 11. You Just Didn't Care Enough

Quality of Will, Causation, and Blameworthiness  
for Actions, Omissions, and Outcomes<sup>1</sup>

(Paper 3)

**Mattias Gunnemyr and Caroline Touborg**

**Abstract.** We refine the intuitively appealing idea that you are blameworthy for something if it happened because you did not care enough. More formally: you are blameworthy for  $X$  (where  $X$  may be an action, omission or outcome) just in case there is the right causal-explanatory relation between your poor quality of will and  $X$ . First, we argue that blameworthiness for actions (etc.) is concerned with *negative* differences: you are blameworthy for the fact that  $X$  occurred instead of  $X^*$ , where  $X$  is worse than  $X^*$ . Second, we argue that the *way* in which your quality of will is poor has to fit *what* you are blameworthy for. With these refinements, the account already gives intuitively correct verdicts in cases of forgetting, making a negative difference to a nevertheless good result, or doing an action with runaway consequences. We then discuss what the right causal-explanatory relation is, and suggest that it is simply causation, understood in the right way. Here, we draw on the account of causation developed by Touborg (2018). According to this account, there are two necessary and jointly sufficient conditions for causation. Roughly,  $C$  causes  $E$  rather than  $E^*$  iff (a)  $C$  is process-connected to  $E$  and (b)  $C$  makes  $E$  more secure and  $E^*$  less secure. With this account of causation, our account of blameworthiness now also gives correct verdicts in omission, pre-emption, and switching cases, Frankfurt-style cases, and collective harm cases.

---

<sup>1</sup> Unpublished manuscript.



# 1. Introduction

Consider the following case, where our immediate reaction is that Suzy is blameworthy for breaking the window:

SOLO SUZY: Suzy is walking down the street. When she reaches the big house on the corner, she stops and considers. She has an intense dislike for the elderly couple who live in the house, and she has just had an idea: she is going to upset them by breaking their window on the first floor. She carefully selects a stone and hurls it towards the window. She feels a jolt of satisfaction when she hears the sound of breaking glass. Then she walks on as if nothing has happened.

We may ask two different questions about this case: we may ask what Suzy is blameworthy *for*, and we may ask *how* blameworthy she is.<sup>1</sup> In this paper, we shall focus on the first of these two questions. In virtue of what is Suzy blameworthy for breaking the window in this case? More generally: what distinguishes cases where an agent is blameworthy for something – an action, omission, or outcome – from cases where she is not?<sup>2</sup>

We aim to develop a compatibilist answer to this question. In doing so, we shall draw on two important approaches in the literature: the quality-of-will approach, and the actual sequence approach. The quality-of-will approach is based on Strawson's suggestion that blame is tied to the reactive attitudes, particularly resentment, and that those attitudes in turn respond to an agent's quality of will:

The reactive attitudes I have so far discussed are essentially reactions to the quality of others' wills towards us, as manifested in their behaviour: to their good or ill will or indifference or lack of concern.

(Strawson 2008/1962: 15)

For developments of this idea, see for instance Watson (1987), Wallace (1994), Smith (2005), McKenna (2012), Talbert (2012), Shoemaker (2015), and Björnsson (2017a, 2017b). Proponents of the quality-of-will approach tend to focus on the

---

<sup>1</sup> These two questions correspond closely to Zimmerman's (2002) distinction between the *scope* of blameworthiness (what you are blameworthy *for*) and *degree* of blameworthiness (*how* blameworthy you are). Zimmerman uses this distinction to reject moral luck, by arguing that while luck matters for what you are blameworthy *for*, it does not matter for your *degree* of blameworthiness. In this paper, by contrast, we stay neutral on the question of moral luck, and focus simply on understanding blameworthiness *for*.

<sup>2</sup> A parallel question may be asked about praiseworthiness for actions, omissions, and outcomes. In the following, we set this aside.

question of *how* blameworthy an agent is (the exception is Björnsson). However, as we will argue, quality of will also fits naturally when we are thinking about blameworthiness *for*.

Second, we shall draw on the actual-sequence approach, which takes its inspiration mainly from Frankfurt-style cases. On this approach, what matters in determining whether an agent is blameworthy for an action, omission, or outcome is the actual causal sequence leading up to that action, omission, or outcome. For developments of this idea, see Fischer and Ravizza (1998) and Sartorio (2016).

Björnsson's account (2017a and 2017b) elegantly combines these two approaches. The basic idea of his account is:

BASIC IDEA: You are blameworthy for X – where X may be an action, omission, or outcome – just in case there is a time *t*, such that your poor quality of will at *t* stands in the right causal-explanatory relation to X.

Exactly how to understand “poor quality of will” is a matter of debate. It means something like manifesting ill will, indifference, or lack of concern (Strawson 2008/1962), caring too little (Björnsson 2017a, 2017b), or showing insufficient regard (McKenna 2012). In our formal definitions, we shall refer simply to “poor quality of will”, leaving it open precisely how this should be understood. In our discussions, though, we often adopt Björnsson's proposal and understand quality of will in terms of “care”. The reader is free to substitute her own preferred understanding of quality of will.

In this paper, we present a new way to develop THE BASIC IDEA. First, we argue that it needs to be refined in a number of ways (Section 2). Next, we present an account of the relevant causal-explanatory connection (Section 3), and finalise the account of when you are blameworthy for actions, omissions and outcomes, testing the account on a number of cases (Section 4). Finally, we show that this account also gives the right verdict in Frankfurt-style cases (Section 5), and in collective harm cases (Section 6).

## 2. Developing the Basic Idea

The basic idea already gives the intuitively right verdict in paradigm cases of blameworthiness *for*, such as SOLO SUZY. Here, Suzy has a poor quality of will – she dislikes the elderly couple who live in the house, and wants to break their window in order to upset them. Furthermore, Suzy's poor quality of will just before she throws her rock stands in the right causal-explanatory relation to the breaking of the window: Suzy's poor quality of will causes/explains her throwing the rock

towards the window, and the breaking of the window. Thus, Suzy is blameworthy both for throwing the rock and for the breaking of the window.

In SOLO SUZY, Suzy intentionally breaks the window. This makes it a paradigm case of blameworthiness *for*. In other cases, however, you may be blameworthy for something even though you did not do it or bring it about intentionally.<sup>3</sup> When you have a poor quality of will, you may forget things you should remember, you may fail to notice things, or neglect to consider them. Suppose, for example, that you do not care as you should, and therefore forget your best friend's birthday. In that case, we think that you are blameworthy for forgetting the birthday – even though, of course, you did not do this intentionally (Smith 2005). The basic idea easily captures this: you are blameworthy for forgetting the birthday, because your poor quality of will – your not caring enough – stands in the right causal-explanatory relation to your forgetting. Or suppose that you plan a weekend at the golf course with your colleagues, without even considering to visit your injured daughter at the hospital (McKenna 2012). Here too, the basic idea captures why you are blameworthy: you failed to even consider visiting your daughter because you did not care enough about her.

However, the basic idea needs a number of refinements. In the remainder of this section, we introduce these refinements gradually, motivated by a series of cases.

The first refinement is motivated by the following observation: when we blame someone for something, this seems to imply that what they are blamed for is *bad*. On its own, however, the basic idea delivers the result that you may be blameworthy for a *good* outcome, if it is caused/explained by your poor quality of will. A simple way to fix this is to add a further necessary condition to the basic idea: you are blameworthy for X – an action, omission, or outcome – only if X is bad. However, this is not quite right. First, it is at best difficult, and at worst impossible, to define what it is for something – an action, omission, or outcome – to be bad *tout court*. It seems much easier to make comparative judgements that an action or outcome is worse than some alternative. Second, there are cases where it seems that you can be blameworthy for making a negative difference, even though the outcome that happens does not seem bad as such. Suppose, for example, that Sally and Bob are cooking a chilli together. Bob is careful about following the recipe. Sally, on the other hand, is more attentive to her phone than to her cooking, and fails to put in some of the ingredients. The chilli still turns out good, though not quite as good as

---

<sup>3</sup> Voluntarists about moral responsibility such as Fischer and Ravizza (1998) and Rosen (2015) would not agree. According to them, voluntary control is a precondition on being blameworthy; and you do not have voluntary control over e.g. forgettings. Still, it seems that you are blameworthy for something in such cases. Here, voluntarists typically argue that you are blameworthy for some earlier action or decision that you did have voluntarily control over (given that you also satisfy some epistemic condition), such as failing to add a note in your calendar about your friend's birthday. This is the *tracing* strategy. There are however some problems with the tracing strategy (see e.g. Smith 2015).

it would have been with all the ingredients. In this case, we think it makes sense to say that Sally is blameworthy for the chilli turning out as it did, even though this outcome is not bad: she is blameworthy for the chilli turning out as it did, rather than turning out even better.

Both of these considerations point towards the same solution: that blameworthiness involves a comparative element. Fully spelled out, you are not simply blameworthy for X, where X is some action, omission, or outcome. Rather, you are blameworthy for the occurrence of X rather than X\*, where X is worse than X\*. This yields the following refined version of the basic idea:

#1 BLAMEWORTHINESS FOR: you are blameworthy for X rather than X\* just in case<sup>4</sup>

- (i) X is worse than X\*, and
- (ii) there is a time *t* such that your poor quality of will at *t* stands in the right causal-explanatory relation to X rather than X\*.

This refined version easily handles the cases we have considered so far. However, problems still remain. Consider the following tragic variation of SOLO SUZY:

TRAGEDY: Everything is as in SOLO SUZY up to the point where the window breaks. But the consequences of the window breaking are dire. The husband is so upset at seeing the broken window that he suffers a heart attack and dies. Unable to cope with her husband's sudden death, the wife has a nervous breakdown, and never fully recovers. Her daughter has to abandon a promising artistic career in Australia, and come home to take care of her mother for the next several years. If Suzy had not broken the window, none of this would have happened. Instead, the couple would have continued to live happily together for many years, and their daughter would have been free to pursue her promising artistic career in Australia.

We have no doubt that Suzy is blameworthy for throwing her rock and for breaking the window. But is she also blameworthy for the runaway consequences: the husband's heart attack? the wife's nervous breakdown? or the end of the daughter's promising artistic career? We do not think so. According to #1 BLAMEWORTHINESS

---

<sup>4</sup> On the intended reading, X is an event that actually occurred, while X\* is a merely possible event that is incompatible with X. This is for example the case with the chilli: the chilli turning out as it did is an actual event, while the chilli turning out even better is a merely possible event. Sally is blameworthy for the fact that the chilli turned out as it did, rather than turning out even better. There is an alternative reading where both X and X\* are events that actually occurred. Suppose, for example, that Ben is also involved in the cooking, and botches the desert. If someone were to blame Sally for the failed desert, we might then correct them by saying "Sally is blameworthy for the chilli rather than (being blameworthy for) the desert". The rather-than construction is ambiguous between these two readings. Throughout the following, we intend the first.

FOR, however, she is: Suzy's poor quality of will causes/explains both the breaking of the window, and the series of unfortunate events that follow.

The case shows that there has to be a tighter fit between *what* an agent is blameworthy for and the *way* in which her quality of will is poor. What is the required fit? Here is a suggestion: an agent is blameworthy for X rather than X\* only if her poor quality of will *specifically in relation to X versus X\** stands in the right causal-explanatory relation to X rather than X\*. We may state the modified condition as follows:

#2 BLAMEWORTHINESS FOR: you are blameworthy for X rather than X\* just in case

- (i) X is worse than X\*, and
- (ii) there is at time *t*, such that your poor quality of will at *t in relation to X versus X\** stands in the right causal-explanatory relation to X rather than X\*.

This captures what we need. To start with an easy case, suppose that although Suzy's quality of will is poor in that she wants to see the elderly couple upset, she still cares as she should about more serious outcomes, such as whether the elderly people might die, or suffer a nervous breakdown, just as she still cares as she should about their daughter's artistic career. If she learned what happened next, she would be horrified and exclaim something like this: "it's true that I wanted to upset them, but I never wanted something like this to happen!". If we were to blame her, for instance, for the husband's death in this case, there clearly would not be the right fit between the *way* in which her quality of will was poor, and *what* we blame her for: although her quality of will was poor in relation to the elderly couple's becoming upset, it was *not* poor in relation to the possibility that the husband might die. Thus, condition (ii) fails to be satisfied.

#2 BLAMEWORTHINESS FOR seems to successfully capture why Suzy is blameworthy for throwing her rock, breaking the window, and upsetting the elderly couple, but not for the husband's death, the wife's nervous breakdown, or the end of the daughter's promising artistic career. On further inspection, however, an unexpected difficulty arises: it is not actually clear that we still get the result that Suzy is blameworthy for throwing her rock, or even for breaking the window. Consider Suzy's throwing her rock. On the revised condition, Suzy is blameworthy for throwing her rock rather than not only if there is a time when she has a poor quality of will in relation to throwing her rock rather than not. But as we have told the story so far, we have not said anything to the effect that Suzy has a poor quality of will in relation to throwing her rock rather than not – we have merely said that she has a poor quality of will in relation to the elderly couple's getting upset. In that case, #2 BLAMEWORTHINESS FOR does not entail that Suzy is blameworthy for breaking the window.

Fortunately, there is an easy way to solve this difficulty. Even though Suzy's throwing her rock is not intrinsically worse than her not doing so, Suzy's throwing her rock *is* worse than not throwing in virtue of how her throwing (rather than not) is related to other things – such as the elderly couple becoming upset. And Suzy's quality of will is poor in relation to her throwing in precisely this sense: she does not care as she should about some of the outcomes that make her throwing worse than not throwing. In particular, she does not care as she should about the elderly couple becoming upset. We may capture this as follows:

#3 BLAMEWORTHINESS FOR: you are blameworthy for  $X$  rather than  $X^*$  just in case there is a  $Y$  and  $Y^*$ , such that

- (i)  $X$  is worse than  $X^*$ , at least partly in virtue of  $Y$  being worse than  $Y^*$ , and
- (ii) there is a time  $t$ , such that your poor quality of will at  $t$  in relation to  $Y$  versus  $Y^*$  stands in the right causal-explanatory relation to  $X$  rather than  $X^*$ .

This secures the verdict that Suzy is blameworthy for throwing the rock: (i) Suzy's throwing the rock ( $X$ ) is worse than her not throwing it ( $X^*$ ), at least partly in virtue of the old couple's becoming upset ( $Y$ ) being worse than their not becoming upset ( $Y^*$ ); and (ii) the time  $t$  just before she throws is such that Suzy's poor quality of will at  $t$  in relation to the elderly couple's becoming upset ( $Y$ ) versus not becoming upset ( $Y^*$ ) stands in the right causal-explanatory relation to her throwing the rock ( $X$ ) rather than not ( $X^*$ ). We similarly get the verdict that Suzy is blameworthy for breaking the window.

In cases where you do have a poor quality of will directly in relation to  $X$  versus  $X^*$ , we may set  $X = Y$  and  $X^* = Y^*$ , effectively making #3 BLAMEWORTHINESS FOR equivalent to its previous iteration #2. In such cases, we will say that  $X$  *just is* worse than  $X^*$  (leaving it open in virtue of what  $X$  is worse than  $X^*$ ). Thus, #3 BLAMEWORTHINESS FOR straightforwardly gives the verdict that Suzy is blameworthy for the elderly couple becoming upset rather than not.

### 3. Characterising the Right Causal-explanatory Relation

Until now, we have relied on an intuitive understanding of “the right causal-explanatory relation”. In this section, we suggest that the relevant relation just is *causation*. The success of this kind of suggestion depends critically on the account of causation that is used. In this section, we consider this in detail, and suggest that the account of causation proposed by Touborg (2018) works well with our account of blameworthiness *for*.

According to this account, there are two necessary and jointly sufficient conditions for causation. First, a cause has to *produce* its effect, in the sense that it has to be connected to its effect via a genuine process. Second, the effect has to *depend* on the cause, in the sense that the security of the effect has to depend on the cause.<sup>5</sup> In the following, we first present the condition of production, and then the condition of dependence. Fully spelled out, both conditions are complex; here we only include as much detail as we need to spell out our account of *blameworthiness for*.

For the sake of simplicity, we presuppose that the laws of nature are deterministic. Correspondingly, we assume determinism in the examples we consider below. However, we believe the account could be extended to also apply to causation in worlds with indeterministic laws.

### **Production as Process-connection**

Let us begin with the production condition. The guiding idea behind this condition is that a cause must be connected to its effect via a genuine process. This idea is familiar from the proposal that causation should be understood in terms of physical processes (see e.g. Dowe 2000). In its simplest form, this proposal may be stated as follows:

PHYSICAL PROCESS: C is a cause of E just in case C is connected to E via a physical process.

A physical process is here understood in terms of transfers of physical quantities – mass, energy, etc. To illustrate the idea, consider a paradigm case of causation, such as Suzy’s throwing her rock and breaking the window. Here, there is indeed a physical process connecting Suzy’s throw, through the trajectory of the rock and its impact on the window pane, to the shattering of the window.

However, trouble is not far to seek: the proposal that a cause must be connected to its effect via a physical process cannot accommodate omissions and absences as causes and effects. This means, for example, that it cannot deliver the intuitively correct verdict on a case like the following:

INDIFFERENT JOHN: John is walking along a beach, and sees a child struggling in the water. John believes that he could save the child with very little effort, and in fact he could, but he is disinclined to expend any energy to help anyone else. He decides not to save the child, and he continues to walk along the beach.

---

<sup>5</sup> This account of causation is inspired by Hall’s (2004) proposal that there are two concepts of causation: the concept of production, and the concept of dependence.

Intuitively, John's failure to jump in the water and save the child is a cause of the child's death. However, John's failure to intervene does not transfer any physical quantities or exert any push or pull on the drowning child; it is a mere absence. Thus, PHYSICAL PROCESS delivers the verdict that John's failure to jump into the water and save the child is *not* a cause of the child's death. This verdict is counterintuitive, and especially so in the context of blame.

The trouble extends further: PHYSICAL PROCESS always delivers the verdict that there is no causal connection when an omission or absence features as an intermediary. Thus, proponents of PHYSICAL PROCESS have to deny that pulling the trigger causes gunshot wounds, or that decapitation causes death, since there is an intermediary absence or omission in both cases: squeezing the trigger removes an obstacle that would have prevented the flight of the bullet; decapitation stops the blood flow, which would have prevented brain starvation. (Such cases are called "double prevention cases", since in these cases C causes E by preventing D, which would have prevented E.)<sup>6</sup>

These cases show that it cannot be a necessary condition for causation that a cause must be connected to its effect via a physical process, when this is understood in terms of transfers of physical quantities. To capture the intuitive idea that some kind of connecting process is necessary for causation, we instead need a more abstract notion of a process, which can include omissions and absences. Touborg suggests that we may get such a more abstract notion of a process by starting from *minimal sufficiency*. Minimal sufficiency is a relation between a set of simultaneous events S and a later event E, where events are understood broadly, so as to include omissions and absences. A set of simultaneous events S is minimally sufficient for a later event E, just in case the occurrence of all the events in S guarantees (given the laws of nature) that E will occur; and if any event is removed from S, the remaining events no longer guarantee (given the laws of nature) that E will occur. Importantly, *minimal sufficiency* is a relation between actual events: only actual events – including actual omissions and absences – may feature in the set S; and the later event E also has to be an actual event (where this includes actual omissions and absences).

Let us say that there is an *apparent process* from C to E when C is connected to E via a *chain* of such relations of minimal sufficiency. This is so when C belongs to a set of simultaneous events  $S_0$ , which is minimally sufficient for a later event  $D_1$ ;  $D_1$  belongs to a set of simultaneous events  $S_1$ , which is minimally sufficient for a later event  $D_2$ ; ... and  $D_n$  belongs to a set of simultaneous events  $S_n$ , which is minimally sufficient for the later event E. When we look more closely – by considering more and more intermediate times between C and E – we may sometimes find that the apparent process from C to E was not genuine: when we consider these intermediate

---

<sup>6</sup> See e.g. Schaffer (2000).



times, we can no longer find a chain of relations of minimal sufficiency connecting C to E. In order for C to be *process-connected* to E, the connection must remain *when we consider more and more intermediate times* between C and E.<sup>7</sup>

This notion of process-connection is sufficiently abstract to accommodate omissions and absences. Returning to the case of INDIFFERENT JOHN, for example, we find that John's poor quality of will (in relation to the child's drowning versus surviving) is process-connected to the child's drowning. John's poor quality of will at the time  $t$  just before he decides not to intervene belongs to a set of simultaneous events that is minimally sufficient for the child's drowning. And this connection remains no matter how many intermediate times we consider. Consider, for example, the intermediate time  $t'$ , after John has decided not to intervene, and before the child has drowned. Here, we find that John's poor quality of will at  $t$  belongs to a set of simultaneous events that is minimally sufficient for his failure to intervene at  $t'$  (remember, his failure to intervene is an actual event), and his failure to intervene at  $t'$  in turn belongs to a set of simultaneous events that is minimally sufficient for the child's drowning. Thus, John's poor quality of will at  $t$  is process-connected to the child's drowning.

The notion of process-connection also allows us to distinguish genuine causes from pre-empted backups in cases such as the following:

BACKUP BILLY: Everything is as in SOLO SUZY, except that Billy also wants the window to break. On seeing that Suzy throws her rock, Billy is satisfied and walks away. However, if Suzy had not thrown her rock, Billy would have thrown a rock himself a moment later, and the window would still have broken.

In this case, while Suzy's and Billy's poor quality of will each guarantees that the window will break, only Suzy's poor quality of will (in relation to the elderly couple's becoming upset versus not) is process-connected to the shattering of the window. To see this, the key is to look at intermediate times. Let  $t$  be the time just before Suzy throws her rock. Then Suzy's poor quality of will at  $t$  belongs to a set of simultaneous events that is minimally sufficient for the shattering of the window; and similarly, Billy's poor quality of will at  $t$  belongs to a set of simultaneous events that is minimally sufficient for the shattering of the window. However, when we bring in more and more intermediate times, we find that we can keep filling in the details in the chain connecting Suzy's poor quality of will to the shattering of the window – going from Suzy's poor quality of will, to her decision to throw, to her throwing the rock, to the rock's trajectory and impact on the window pane. By contrast, the connection between Billy's poor quality of will at  $t$  and the shattering

---

<sup>7</sup> The full definition of process-connection includes a further refinement: to be able to handle all cases of late pre-emption, it makes use of a more demanding relation of minimal sufficiency, namely *time-sensitive sufficiency*. For simplicity, we leave out this refinement.

of the window breaks down when we consider intermediate times. Consider, for example, a time  $t'$  after Suzy has thrown her rock and Billy has turned away, but before the window shatters. To connect Billy's poor quality of will to the breaking of the window, we would need an event D at this time  $t'$  – such as Billy's rock flying towards the window – so that Billy's poor quality of will belongs to a set of events that is minimally sufficient for D, and D in turn belongs to a set of events that is minimally sufficient for the window-shattering. But there is no such event D in the actual world. For this reason, Billy's poor quality of will at  $t$  is not process-connected to the breaking of the window. This fits the judgement that Suzy's poor quality of will at  $t$  is a cause of the shattering of the window, while Billy's poor quality of will is not. In this way, the notion of process-connection does crucial work in distinguishing genuine causes from pre-empted backups.

However, process-connection is not sufficient for causation. The condition of process-connection needs to be supplemented by a second necessary condition for causation, requiring that a cause must make a difference to its effect. The need for this is brought out by the following three considerations.

First, process-connection on its own cannot yield the intuitively correct verdict on counterexamples to the transitivity of causation, such as the following switching case:

TROLLEY TROUBLE: Suzy is standing by a switch in the tracks as a trolley approaches in the distance. If she flips the switch, the trolley will travel down the left-hand track; if she does not flip the switch, it will travel down the right-hand track. Further ahead, the tracks converge again, and after that, five people are tied to the then single track. Suzy wants the five to get run over, and she erroneously believes that they are tied to the left-hand track. She flips the switch so that the trolley travels down the left-hand track, and subsequently runs over the five people. However, if she had not flipped the switch, the trolley would still have run over the five, reaching them via the right-hand track.<sup>8</sup>

Intuitively, Suzy's poor quality of will at  $t$  (the time just before she flips the switch) is not a cause of the five getting run over. However, Suzy's poor quality of will at  $t$  is process-connected to the five getting run over. Thus, we cannot simply understand causation in terms of process-connection.

You might not immediately notice that there is a process-connection in this case since Suzy's poor quality of will at  $t$  does not belong to a set of simultaneous events that is itself minimally sufficient for the five getting run over: the set just containing the approach of the trolley, and the layout of the tracks, and so on, is sufficient for the death of the five. However, there is a *chain* connecting Suzy's poor quality of

---

<sup>8</sup> This case is inspired by Foot (1967), Thomson (1976) and Van Inwagen (1978). Switching cases like this are also common in the causation literature. See Hall (2007), and Paul and Hall (2013).

will at  $t$  to the death of the five: Suzy's poor quality of will at  $t$  belongs to a set of simultaneous events that is minimally sufficient for the trolley's journey along the left-hand track, and the trolley's journey along the left-hand track belongs to a set of simultaneous events that is minimally sufficient for the five getting run over. And this connection remains when we consider more intermediate times.

The problem arises since process-connection is a transitive relation – if  $C$  is process-connected to  $D$ , and  $D$  is process-connected to  $E$ , then  $C$  is process-connected to  $E$ . By contrast, causation is not transitive: it may happen that  $C$  is a cause of  $D$ , and  $D$  is a cause of  $E$ , but  $C$  is not a cause of  $E$  – as in TROLLEY TROUBLE.

Second, process-connection cannot on its own accommodate *contrastive* causal claims. Process-connection is simply a relation between two actual events: an actual event  $C$  is process-connected to an actual event  $E$ . However, contrastive causal claims include counterfactual contrasts to the cause  $C$  or the effect  $E$ , and the truth-value of a contrastive claim depends on what these contrasts are. The need to handle such contrastive causal claims is especially pressing when we are concerned with blameworthiness for actions, omissions, and outcomes: as we have seen above, BLAMEWORTHINESS FOR is based precisely on a contrastive claim, namely that your poor quality of will stands in the right causal-explanatory relation to  $X$  *rather than*  $X^*$ .

Third, process-connection on its own cannot distinguish between causes and background conditions. Suppose, for example, that Selma has no royal connections. Is the queen of Sweden's failure to water Selma's flowers a cause of their death? Intuitively, it is not.<sup>9</sup> However, the queen's failure to water Selma's flowers *is* process-connected to their death. So if we take process-connection to be sufficient for causation, we cannot accommodate the intuitive verdict in this case.

These difficulties have a common solution: recognising that there is a second necessary condition for causation, which captures the intuitive idea that a cause must make a relevant difference to its effect.

## Dependence and Security

The core idea that causes are difference-makers is familiar. For example, David Lewis writes that “[w]e think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it” (Lewis 1973a: 557). This idea is the starting point for counterfactual accounts of causation. In its contrastive form, it may be stated as follows:

---

<sup>9</sup> See e.g. Hart and Honoré (1985) or Sartorio (2004).

SIMPLE: Suppose that C occurs at  $t$ , E occurs later, and E\* is incompatible with E.

Then C is a cause of E rather than E\* just in case  
if C had not occurred, then E\* would have occurred instead of E.

The heart of SIMPLE is the counterfactual: “if C had not occurred, then ...”. To evaluate this counterfactual, we first identify all the worlds where C does not occur. Among these, we consider the worlds that are *closest* to the actual world @. If the consequent is true in each of these worlds, then the counterfactual is true; otherwise it is false.

The relevant notion of closeness is standardly understood in terms of similarity between entire worlds. Following Paul and Hall (2013), we prefer instead to understand it in terms of similarity between states of worlds at times. Thus, we shall say that two possible worlds  $w$  and  $w^*$  are close-at-time- $t$  to the extent that the state of  $w$  at  $t$  is similar to the state of  $w^*$  at  $t$ . Supposing that C occurs at time  $t$ , this means that the counterfactual “if C had not occurred, then ...” is true just in case the consequent is true in each of the closest-to-@-at- $t$  worlds where C does not occur.

Even with this clarification, a question remains: what replaces C in the closest-to-@-at- $t$  worlds where C does not occur? An obvious answer is that C is replaced by an event that is as similar as possible to C, without satisfying C’s conditions of occurrence. However, this proposal yields intuitively false results. As Lewis writes: “if C had not occurred and almost-C had occurred instead, very likely the effects of almost-C would have been much the same as the actual effects of C. So our causal counterfactual will not mean what we thought it meant, and it may well not have the truth-value we thought it had” Lewis (2004: 90). That will not do. We need an alternative proposal about what replaces C.

Our preferred answer is that, when we evaluate counterfactuals, we do not in fact consider *all* possible worlds. Rather, we only consider a restricted class of possible worlds – namely, those possible worlds that we take to be relevant. This restricted class of possible worlds is itself a causal relatum; we shall call it a *possibility horizon*. Our chosen possibility horizon will typically not contain any worlds where C is replaced by almost-C. Rather, it will typically contain only worlds where C either occurs or is replaced by a contextually salient alternative C\* that is quite different from C. When we only consider the worlds within such a possibility horizon, we find that in the *closest* worlds where C does not occur, it is replaced by C\*.

Based on this, we may now give the following more developed version of SIMPLE:

SIMPLE\*: Suppose that C occurs at  $t$ , E occurs later, and E\* is incompatible with E.

Then C is a cause of E rather than E\* within possibility horizon H just in case  
there is at least one world in H where C does not occur, and in the closest-to-@-at- $t$   
world(s) in H where C does not occur, E\* occurs instead of E.

SIMPLE\* can capture our intuitions in a wide range of cases. In particular, it successfully handles the cases that presented difficulties for process-connection. In TROLLEY TROUBLE, SIMPLE\* entails that Suzy's poor quality of will is not a cause of the five getting run over, because Suzy's poor quality of will makes no difference to their fate – the five would have died either way. Furthermore, SIMPLE\* is tailor-made to handle contrastive causal claims, such as “Sally's not caring about the chilli caused the chilli to be just good, rather than excellent”. Finally, SIMPLE\* captures the verdict that the queen of Sweden's failure to water Selma's flowers did not cause them to die: in ordinary contexts, it is not a relevant possibility that the queen waters Selma's flowers. Rather, the queen's failure to water the flowers is treated as a background condition. Thus, the possibility horizon that is in play in an ordinary context only contains worlds where the queen does not water the flowers.

As is well known, however, SIMPLE\* does not give a necessary condition for causation: there are cases where C is clearly a cause of E, even though E would still have occurred if C had not. We have already seen such a case: in BACKUP BILLY, it is clear that Suzy's throwing her rock is a cause of the window shattering. However, if Suzy had not thrown her rock, the window would still have shattered – because, in that case, Billy would have thrown *his* rock. Cases such as this show that in order to capture the idea that making a difference is necessary for causation, we need a more subtle notion of difference-making: one that can capture, for instance, how Suzy's throw makes a difference to the shattering of the window, even though the window would still have shattered if she had not thrown.

The key to developing such a more subtle notion of difference-making is to pay attention to the *modal* features of events. In particular, when an event actually occurs, we may ask how easily it could have failed to occur; and when an event does not occur, we may ask how easily it *could* have occurred.<sup>10</sup> Touborg (2018) uses the notion of *security* to capture this:

Whenever an effect actually occurs, it has *positive security*. However, it may have a higher or lower *degree* of positive security. In some cases, an event E actually occurs, but when we consider what was the case at some earlier time *t*, we find that if things had been just slightly different at time *t*, E would not have occurred. In such cases, we shall say that E had a low degree of positive security at time *t*. Suppose, for example, that Suzy in SOLO SUZY throws her rock towards the window and breaks it, but if there had been just a slight gust of wind at *t* (the time when Suzy threw her rock), a swaying branch would have deflected her rock, and the window would have remained intact. In this case, the breaking of the window has a very low degree of positive security at *t*. In other cases, an event E actually occurs, and when

---

<sup>10</sup> More carefully: when a particular type of event does not occur, we may ask how easily an event of this type could have occurred. Speaking of types of events solves the difficulty that we cannot refer determinately to an event that did not occur. For simplicity, we suppress this complication in the text.

we consider what was the case at some earlier time  $t$ , we find that things would have had to be quite different at  $t$  in order for E not to occur. In such cases we shall say that E had a high degree of positive security at  $t$ .

Whenever an event fails to occur, this event has *negative security*. Once again, it may then have a higher or lower *degree* of negative security. Consider some event E that does not actually occur, and consider some time  $t$  prior to the time when E would have occurred, if it did occur. We may now ask: how different would things have to be at  $t$  in order for E to occur? If things would only have to be ever so slightly different at  $t$  in order for E to occur, we shall say that E has a low degree of negative security at  $t$ : although E does not happen, circumstances at  $t$  are such that it is *close* to happening. If, on the other hand, things would have to be quite different at  $t$  in order for E to happen, we shall say that E has a high degree of negative security: considering the circumstances at  $t$ , E is *far* from happening.

More formally, we may understand security-at-a-time in terms of the distance-at-a-time between worlds. We have already introduced the notion of distance-at-a-time above, when we discussed the evaluation of counterfactuals. As a reminder: two possible worlds  $w$  and  $w^*$  are close-at-time- $t$  to the extent that the state of  $w$  at  $t$  is similar to the state of  $w^*$  at  $t$ . Based on this, we may define security-at-a-time as follows:

If an event E occurs in  $w$ , then E has positive security in  $w$ , and its degree of positive security at an earlier time  $t$  is given by the distance-at- $t$  between  $w$  and the closest-to- $w$ -at- $t$  world(s) where E does not occur.

If an event E does not occur in  $w$ , then E has negative security in  $w$ , and its degree of negative security is given by the distance-at- $t$  between  $w$  and the closest-to- $w$ -at- $t$  world(s) where E occurs.

This notion of security allows us to capture a more subtle notion of difference-making: making a difference to the *security* of an event. A cause does not have to make a difference as to whether its effect occurs or not. But it *does* have to make a difference to the security of its effect: supposing that C occurs at time  $t$ , it has to be the case that if C had not occurred, E would have been *less secure* at  $t$  than it actually was. In the case of contrastive causal claims, such as “C is a cause of E rather than E\*”, C has to make a difference to the security of both E and E\*: supposing again that C occurs at  $t$ , it has to be the case that if C had not occurred, E would have been *less secure* at  $t$  and E\* would have been *more secure* at  $t$  than what was actually the case.<sup>11</sup>

---

<sup>11</sup> The suggestion that a cause must make its effect more secure is somewhat similar to the controversial suggestion that a cause must raise the probability of its effect. In particular, (apparent) counterexamples to the suggestion that causes are probability-raisers can be translated

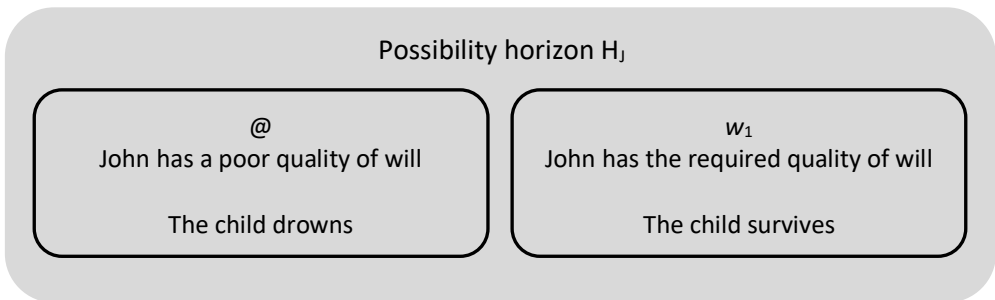
## Causation

So far, we have introduced two necessary conditions for causation: the condition of process-connection, and the condition of difference-making, where difference-making is understood in terms of security. Neither of these two conditions can stand alone. The condition of process-connection needs help from security when dealing with switching cases, contrastive causal claims, and the distinction between causes and background conditions; the security-condition needs help from process-connection when dealing with pre-emption cases such as *BACKUP BILLY*. But together, these two conditions are jointly sufficient for causation, yielding the following account:<sup>12</sup>

CAUSATION: Suppose that C occurs at  $t$  and E occurs later.  
Then C is a cause of E rather than E\* within possibility horizon H just in case

- (a) C is process-connected to E,
- (b) there is at least one world in H where C does not occur,  
and in the closest-to-@-at- $t$  world(s) in H where C does not occur,  
E is *less secure* at  $t$  and E\* is *more secure* at  $t$  than they are in @.

This account of causation handles the cases we have considered so far. Consider first *INDIFFERENT JOHN*. As before, let  $t$  be the time just before John decides not to intervene. We have already seen that John's poor quality of will at  $t$  is process-connected to the child's drowning. We may now consider whether John's poor quality of will also satisfies the condition of difference-making within the possibility horizon below:



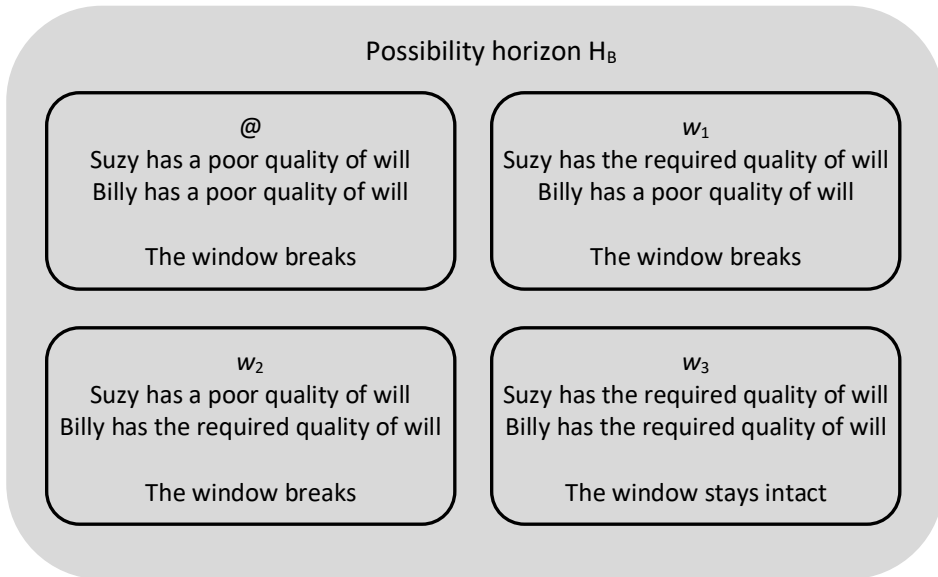
---

into (apparent) counterexamples to the suggestion that causes make their effects more secure. However, the notion of security within a possibility horizon offers resources to resist such counterexamples. We therefore do not think they threaten the proposal. For reasons of space, however, we do not discuss such examples further.

<sup>12</sup> Note that we include effect-contrasts in this definition.

Within  $H_J$ , the closest-to-@-at- $t$  world where John does not have a poor quality of will at  $t$  is  $w_1$ , where he has the required quality of will at  $t$ . Here, John jumps into the water and saves the child. Thus, the child's drowning has positive security in @ at  $t$  (since it occurs in @) and negative security in  $w_1$  (since it does not occur in  $w_1$ ). From this, it immediately follows that the child's drowning is *less secure* at  $t$  in  $w_1$  than it is in @. Similarly, the child's survival is *more secure* at  $t$  in  $w_1$  than it is in @. Thus, CAUSATION yields the result that John's poor quality of will at  $t$  is a cause (within  $H_J$ ) of the child's drowning rather than surviving. Therefore, John is blameworthy for the child's drowning rather than surviving.<sup>13</sup>

Consider next BACKUP BILLY. As before, let  $t$  be the time just before Suzy throws her rock. We may then consider what caused the window-shattering, within the following possibility horizon:



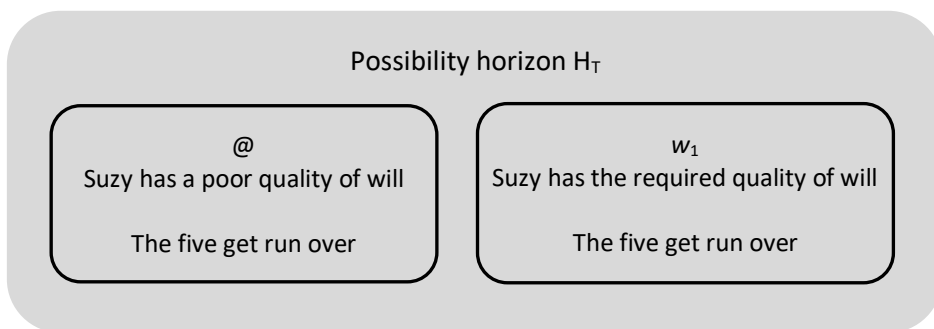
The possibility horizon  $H_B$  includes a salient alternative to Suzy's poor quality of will (in relation to the elderly couple's becoming upset versus not) – namely, her

<sup>13</sup> Is John also blameworthy for killing the child? Killing is sometimes understood simply as causing someone's death. If killing is understood in this way, then we would have to say that John killed the child. However, we think there are further conditions on killing (roughly related to the distinction between doing and allowing, see e.g. Woollard 2015), and John does not satisfy those further conditions – he merely allows the death of the child.



having the required quality of will; and it includes a salient alternative to Billy's poor quality of will – namely, his having the required quality of will. Independently of this choice of possibility horizon, we have already seen that Suzy's poor quality of will at time  $t$  is process-connected to the shattering of the window, while Billy's poor quality of will is not. We may now verify that Suzy's poor quality of will at  $t$  makes a difference to the security of the window-shattering. The closest-to-@-at- $t$  world within  $H_B$  where Suzy does not have a poor quality of will at  $t$  is  $w_1$ , where she has the required quality of will at  $t$ . The window still shatters in  $w_1$ , since Billy has a poor quality of will and therefore throws his rock when Suzy does not. However, the window-shattering is *less secure* at  $t$  in  $w_1$  than it is in @: compared with @,  $w_1$  is *closer-at- $t$*  to  $w_3$  where the window does not shatter. In  $w_1$  only one thing needs to change at  $t$  in order for the window not to break (namely Billy's poor quality of will), whereas in the actual world @, two things need to change at  $t$  in order for the window not to break. And similarly, the window's remaining intact is *more secure* at  $t$  in  $w_1$  than it is in @. Thus, CAUSATION yields the desired result: Suzy's poor quality of will at  $t$  is a cause (within possibility horizon  $H_B$ ) of the window shattering rather than remaining intact, while Billy's poor quality of will is not.

Finally, consider TROLLEY TROUBLE. We have already seen that Suzy's poor quality of will at  $t$  (the time just before she flips the switch) is process-connected to the five getting run over. Now consider the possibility horizon  $H_T$  below, where the relevant alternative to Suzy's having a poor quality of will at  $t$  (in relation to the five) is her having the minimally required quality of will:



Within this possibility horizon, Suzy's poor quality of will at  $t$  does not make any difference to the *security* of the five getting run over: there *is* no world where the five are not run over. Thus, their getting run over is infinitely secure, both in @ and in  $w_1$ . And so, their getting run over is *just as secure* in  $w_1$  as it is in @. We therefore

find, as we should, that Suzy's poor quality of will is not a cause (within possibility horizon  $H_T$ ) of the five getting run over rather than not.

## 4. Completing the Account

We suggest that the causal-explanatory relation that has to hold between an agent's poor quality of will and what she is blameworthy for is *causation*, understood as suggested above.

As we have seen, causation is relativised to a possibility horizon. Thus, it may sometimes be the case that  $C$  is a cause of  $E$  rather than  $E^*$  within possibility horizon  $H_1$ , while  $C$  is not a cause of  $E$  rather than  $E^*$  within a different possibility horizon  $H_2$ . This feature of the general account of causation has a number of advantages. However, it would be unsatisfactory to say, for instance, that you are blameworthy for  $X$  rather than  $X^*$  within possibility horizon  $H_1$ , but not within possibility horizon  $H_2$ . We may avoid this relativity by insisting that what matters for blameworthiness is causation within the *relevant* possibility horizon. This raises a crucial question: what is the relevant possibility horizon when evaluating what an agent is blameworthy for?

Suppose we are evaluating whether your poor quality of will at time  $t$  (in relation to  $Y$  versus  $Y^*$ ) is a cause of  $X$  rather than  $X^*$ , and that the purpose of this evaluation is to determine whether you are blameworthy for  $X$  rather than  $X^*$ . To make this evaluation, we start from the actual state of the world at time  $t$ . We then identify relevant alternatives to the way things were at time  $t$ . If you had a poor quality of will at  $t$  (in relation to  $Y$  versus  $Y^*$ ), we think it is relevant that you could instead have had the quality of will (in relation to  $Y$  versus  $Y^*$ ) that you were minimally required to have.<sup>14</sup> By contrast, it is not relevant that you could have had an even worse quality of will, or a saintly quality of will far above what was minimally required. Similarly, if someone else had a poor quality of will at  $t$ , we think it is relevant that *they* could have had the quality of will they were minimally required to have. But again, it is not a relevant possibility that they could have had an even worse quality of will, or a saintly quality of will. Other changes to what actually happened at  $t$  may or may not be relevant as well. This gives a criterion for determining which possible worlds belong to the relevant possibility horizon: if a possible world  $w$  represents a relevant alternative to how things were at time  $t$ , and evolves forward in accordance with the laws of nature, then it is included in the

---

<sup>14</sup> This proposal is closely related to Björnsson's (2017a and 2017b) proposal that what matters is how your quality of will falls short of what could be demanded.

relevant possibility horizon. Otherwise not. We may summarise this in the following rule of thumb:<sup>15</sup>

RELEVANT POSSIBILITIES FOR BLAME: To determine, for the purpose of attributing blame, whether your poor quality of will at time  $t$  (in relation to  $Y$  versus  $Y^*$ ) is a cause of a later event  $X$  rather than  $X^*$ , it is a relevant possibility that you could instead have had the minimally required quality of will at  $t$  (in relation to  $Y$  versus  $Y^*$ ). Similarly, it is a relevant possibility that anyone else involved in the situation who had a poor quality of will at time  $t$  could have had the minimally required quality of will at time  $t$ . Every combination of these possibilities is relevant. Other possibilities may or may not be relevant as well.<sup>16</sup>

When we discuss collective harm cases in Section 6, we motivate why we have to include the possibility that each agent involved in the situation could have had the required quality of will.<sup>17</sup>

With this, we may now complete our account of blameworthiness for actions, omissions, and outcomes as follows:

BLAMEWORTHINESS FOR: you are blameworthy for  $X$  rather than  $X^*$  just in case there is a  $Y$  and  $Y^*$ , such that

- (i)  $X$  is worse than  $X^*$ , at least partly in virtue of  $Y$  being worse than  $Y^*$ , and
- (ii) there is a time  $t$ , such that your poor quality of will at  $t$  in relation to  $Y$  versus  $Y^*$  is a cause of  $X$  rather than  $X^*$ , within the relevant possibility horizon  $H$ .<sup>18</sup>

---

<sup>15</sup> What matters here is simply that these possibilities are relevant alternatives to the state of the actual world at time  $t$ . It does not matter whether the actual world could, from an earlier state, evolve into one of these alternative states (given determinism, it of course could not).

<sup>16</sup> This principle presupposes that everyone involved is an agent. If someone is not an agent at the relevant time – if he or she for instance is insane, under the influence of drugs, etc. – this person may be exempt from the requirement that they should have a particular quality of will. In that case, the relevant possibility horizon for determining blameworthiness may not include the possibility that they could have had a different quality of will. As a result, our account will yield the result that they are not blameworthy. In the following, we set such cases aside.

<sup>17</sup> You might wonder how to decide which agents are involved in a situation. Here, it is important to note that RELEVANT POSSIBILITIES FOR BLAME is meant to ensure that enough agents are included. The verdict of our account will not change if even more agents are included. If you are unsure whether an agent is involved, the rule of thumb is to include him.

<sup>18</sup> We here set aside cases of deviant causation. To handle such cases, something further is needed – at the very least, a requirement that your bad quality of will is a *non-deviant* cause of  $X$  rather than  $X^*$ . However, since cases of deviant causation present trouble for everyone, we think it is appropriate to set them aside for now.

We will now test this account on the cases we have considered so far.

Consider first INDIFFERENT JOHN. In this case, (i) the child's drowning *just is* worse than its surviving. Furthermore, (ii) John has a poor quality of will at *t* (the time just before he decides not to intervene) in relation to the child's drowning versus surviving. And as we have already seen, John's poor quality of will at *t* in relation to the child's drowning versus surviving is a cause of the child's drowning rather than surviving, within the relevant possibility horizon.

Consider next BACKUP BILLY. This case is just like SOLO SUZY, except that Billy is lurking in the background. There is a *Y* and *Y\** – the elderly couple's becoming upset versus not becoming upset – such that (i) the breaking of the window (*X*) is worse than its staying intact (*X\**), at least partly in virtue of the elderly couple's becoming upset (*Y*) being worse than their not becoming upset (*Y\**). Furthermore, we have already seen that (ii) Suzy's poor quality of will at *t* (the time just before she throws her rock) in relation to the elderly couple's becoming upset (*Y*) versus not (*Y\**) is a cause of the window breaking (*X*) rather than staying intact (*X\**), within the relevant possibility horizon. Thus, BLAMEWORTHINESS FOR yields the result that Suzy is blameworthy for the window breaking rather than staying intact. By contrast, Billy is not blameworthy for this, since there is no process connecting Billy's poor quality of will to the breaking of the window.

Finally, consider TROLLEY TROUBLE. Here, (i) the five getting run over *just is* worse than their not getting run over. Furthermore, (ii) Suzy has a poor quality of will at *t* (the time just before she flips the switch to the left) in relation to the five getting run over. However, as we have seen, Suzy's poor quality of will at *t* is *not* a cause of the five getting run over rather than not within the relevant possibility horizon, since it does not make any difference to the security of their getting run over. Thus, BLAMEWORTHINESS FOR delivers the intuitively correct result: Suzy is not blameworthy for the five getting run over rather than not.<sup>19</sup>

## 5. Blameworthiness and Frankfurt-style Cases

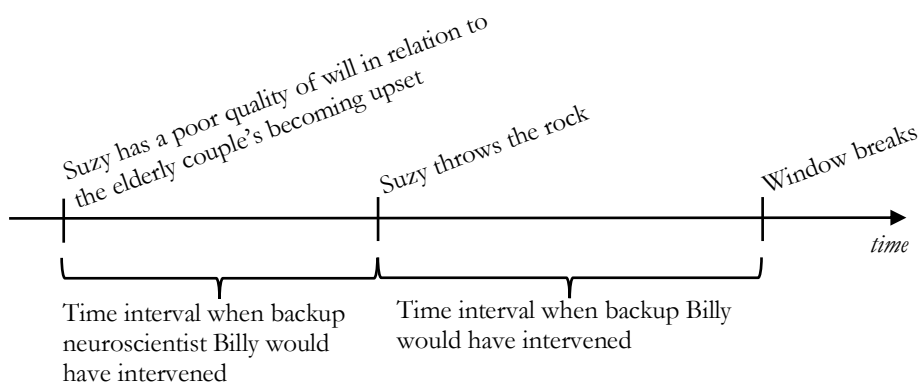
Frankfurt-style cases, first introduced by Frankfurt (1969), are important test cases for accounts of blameworthiness. In these cases, there is a backup ensuring that an agent will act in a certain way. However, as things turn out, the backup does not have to intervene. For example, consider the following Frankfurt-style variation of SOLO SUZY:

---

<sup>19</sup> This is of course consistent with Suzy's being blameworthy for other things – for example, for intending to kill the five. We may wonder whether it makes any difference for *how* blameworthy Suzy is that – because of circumstances outside of her control – she does not actually kill the five. This is related to the question of moral luck, which we have set aside in this paper.

BACKUP NEUROSCIENTIST BILLY: everything is as in SOLO SUZY, except for the following: unbeknownst to Suzy, the mischievous neuroscientist Billy has implanted a chip in Suzy's brain and is now monitoring her process of deliberation. Billy wants Suzy to throw a rock and break the window. If he thinks, as a result of his monitoring of Suzy's process of deliberation, that Suzy is not going to throw a rock and break the window, he will have the chip induce this intention in her and make her act on it. As things happen, Billy does nothing.

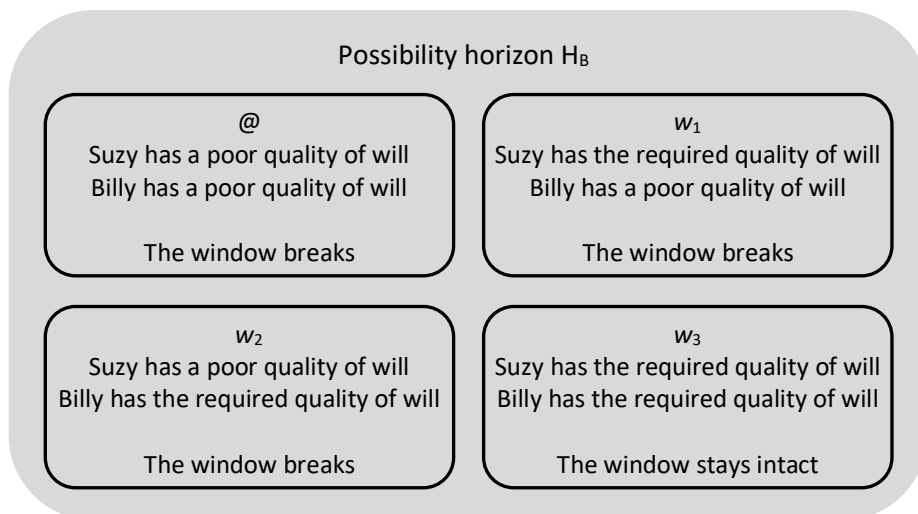
Frankfurt-style cases such as BACKUP NEUROSCIENTIST BILLY have a similar structure to early pre-emption cases such as BACKUP BILLY, where Billy would have thrown a rock at the window had Suzy not done so. In BACKUP BILLY as well as in BACKUP NEUROSCIENTIST BILLY, Billy is lurking in the background with sinister intent. The difference is that in one case, Billy is going to intervene by modifying Suzy's intention, whereas in the other he is going to intervene by breaking the window himself. If we consider the process connecting Suzy's poor quality of will to the breaking of the window, we see three salient features. First, there is her poor quality of will, then there is her throwing of the rock, and finally there is the breaking of the window.



BLAMEWORTHINESS FOR will give the same result no matter where in the process connecting Suzy's poor quality of will to the window breaking Billy is ready to enter as a back-up. This way, BLAMEWORTHINESS FOR straightforwardly gives the intuitively right answer in Frankfurt-style cases. In BACKUP NEUROSCIENTIST BILLY, for instance, BLAMEWORTHINESS FOR gives the verdict that Suzy is blameworthy for breaking the window:

To begin with, (i) the window breaking is worse than its staying intact, at least partly in virtue of the elderly couple's becoming upset being worse than their not becoming upset. Furthermore, (ii) let  $t$  be the time right before Billy would otherwise have

intervened via the chip. At this time, Suzy has a poor quality of will but could have had the required quality of will, and the same goes for Billy. This yields a possibility horizon with four possible worlds, as follows:



Given  $H_B$ , Suzy's poor quality of will at  $t$  in relation to the elderly couple's becoming upset caused the window to break within the relevant possibility horizon. (a) We have already seen that Suzy's poor quality of will at  $t$  (in relation to the elderly couple's becoming upset) is process-connected to the breaking of the window in *BACKUP BILLY*, and precisely the same reasoning shows that it is also process-connected to the breaking of the window here. Moreover, (b) the window breaking is *less secure* at  $t$  in the closest world where Suzy has the minimally required quality of will. When Suzy has the minimally required quality of will, only one thing would need to change at  $t$  in order for the window not to break (namely Billy's poor quality of will), whereas two things would have to change at  $t$  in order for the window not to break in the actual world (namely both Suzy's poor quality of will and Billy's poor quality of will). For similar reasons, the window's staying intact is *more secure* at  $t$  in the closest possible world where Suzy has the required quality of will.

Ever since Frankfurt (1969), cases like *BACKUP NEUROSCIENTIST BILLY* are frequently launched against the idea that blameworthiness requires alternate possibilities – or more broadly, that moral responsibility requires such possibilities:

PRINCIPLE OF ALTERNATE POSSIBILITIES (PAP): a person is morally responsible for what he has done only if he could have done otherwise.<sup>20</sup>

(Frankfurt 1969: 829)

Before Frankfurt (1969), most writers agreed that something like PAP was true. However, incompatibilists and compatibilists disagreed about how to understand the “could” in “could have done otherwise”. Incompatibilists argued that the “could” in PAP requires indeterminism. To use a common metaphor, it requires that there is a fork in the road.<sup>21</sup> If PAP is understood in this way, it entails that moral responsibility is incompatible with determinism. Against this, compatibilists argued that the “could” in PAP does not require indeterminism. According to classical compatibilists such as Hume (in *An Enquiry Concerning Human Understanding*), the agent’s ability to do otherwise should be analysed in conditional terms: you *could* have done otherwise if and only if it is the case that you *would* have done otherwise if you had willed (wanted, chosen or decided) to do otherwise.<sup>22</sup> On this interpretation, PAP states that you are blameworthy for an action only if it is the case that you would have done otherwise if you had had a different will. For instance, Suzy is blameworthy for breaking the window only if she would not have thrown the stone if she had had a different will.<sup>23</sup>

Cases such as BACKUP NEUROSCIENTIST BILLY cast doubt on the classical compatibilist version of PAP: we intuitively judge that Suzy is blameworthy for breaking the window even though (thanks to Billy) she would still have broken the window if she had had the required quality of will at *t*. The problem, we think, is that the classical compatibilist version of PAP imposes a very strong requirement on blameworthiness, namely that for instance Suzy’s poor quality of will should make a *counterfactual difference* to the window breaking. This requirement is stronger than our requirement that for instance Suzy’s poor quality of will should be a *cause* of the window breaking. And the significance of this difference is brought out precisely in cases of pre-emption, such as BACKUP BILLY and BACKUP NEUROSCIENTIST BILLY: in these cases, Suzy’s poor quality of will is a *cause* of the

---

<sup>20</sup> As PAP is stated, it concerns actions and omissions but not outcomes. However, it is usually applied to outcomes as well (See Van Inwagen 1978).

<sup>21</sup> See Chisholm (1964b). Van Inwagen (1978) is a more recent advocate of the incompatibilist position.

<sup>22</sup> See also e.g. (G. E. Moore 1912: ch. 6). For counterexamples to the classical compatibilist counterfactual analysis of “could”, see e.g. Chisholm (1964b) and Van Inwagen (1983).

<sup>23</sup> There are also other interpretations of “could” that we have not addressed here, for instance Wallace’s (1994) suggestion that “could” should be interpreted as concerning general abilities rather than special abilities.

window breaking, but because of Billy, Suzy's poor quality of will does not make a counterfactual difference to the window breaking.

## 6. Blameworthiness in Collective Harm Cases

Finally, our account sheds light on collective harm cases; that is, cases where no single action is necessary or sufficient for bringing about some bad outcome. Consider the following case:

THE LAKE: Ann, Beth, and Claire live close to a lake with a sensitive ecosystem. Each of them has a boat. They have all been using a cheap and hazardous paint. However, they have recently learned that this has brought the ecosystem to the verge of collapsing. Each of them now believes that if she were to switch to a more expensive but environmentally friendly paint, this might be enough to save the ecosystem from collapse. Still, none of the three cares sufficiently about the ecosystem in the lake. All three continue to use the hazardous paint, and the lake becomes a wet wasteland. As a matter of fact, the ecosystem would still have collapsed if just one of them had switched to the environmentally friendly paint. However, if two or more of them had switched to the environmentally friendly paint, the ecosystem would not have collapsed.

(Adapted from Björnsson 2011)<sup>24</sup>

Intuitions about this case might vary. It might seem that Ann, Beth and Claire are blameworthy for the collapse of the ecosystem since the ecosystem collapsed as a result of what they did. It might also seem that none of them is blameworthy since it is true of each that the ecosystem would have collapsed whether or not she had switched to the environmentally friendly paint. If someone for instance blames Ann for the collapse of the ecosystem, she might make the following defence:

DEFENCE: "I accept that I may be blameworthy for using the hazardous paint. But I'm not blameworthy for the collapse of the ecosystem. Given that Beth and Claire didn't care, my lack of care didn't matter."

Moreover, even if we think that Ann, Beth and Claire are blameworthy for the collapse of the ecosystem, there is a puzzle about whether they are to blame individually or collectively. On the one hand, it seems that they are not individually

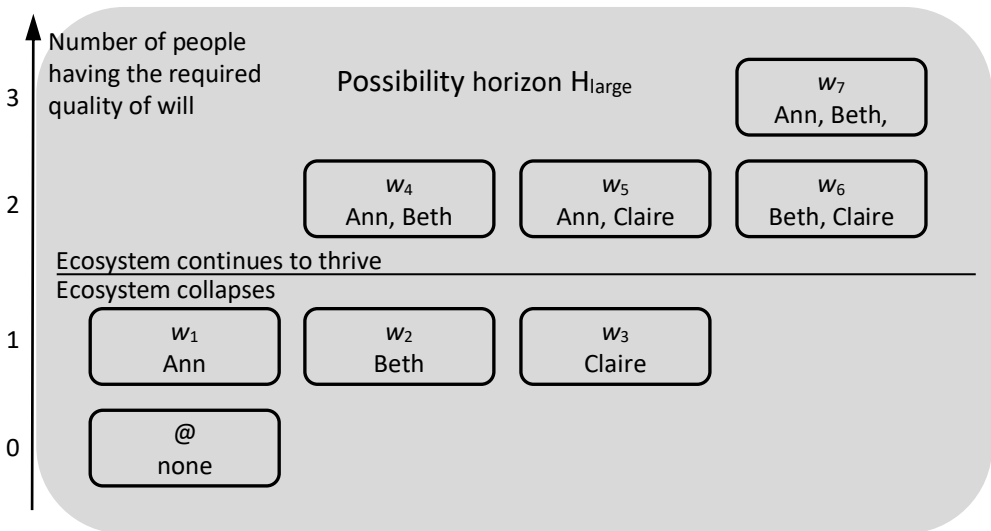
---

<sup>24</sup> In this paper, Touborg and I use a slightly different version of the lake than we did in "Reasons for Action". In this version, it is Ann, Beth and Claire that paints their boats, while in the previous version, it is you, Vanessa and Walter who paint your boats. There are no other significant differences. I hope that this slight change does not confuse things too much.



to blame since each has the defence just sketched readily at hand. On the other hand, it seems that they could not be collectively to blame since they do not constitute a collective. They did not perform an intentional collective action,<sup>25</sup> and they do not share a formal decision procedure.<sup>26</sup>

We suggest that the conflicting intuitions about THE LAKE can be understood as arising from different choices of possibility horizon – where the choice of possibility horizon may, in turn, reflect either a collective or an individual perspective. Consider first the time  $t$  before any of the three has painted their boats. Following the standard procedure for generating the relevant possibility horizon based on what is happening at time  $t$  (as stated in RELEVANT POSSIBILITIES FOR BLAME), we get a possibility horizon that contains every combination of Ann, Beth and Claire having their actual poor quality of will at  $t$ , and having the minimally required quality of will at  $t$ . Call this possibility horizon  $H_{large}$ . In some worlds within this possibility horizon, the ecosystem will be saved (for instance in the world where all three have the required quality of will). In other worlds, the ecosystem will collapse (as in the actual world).



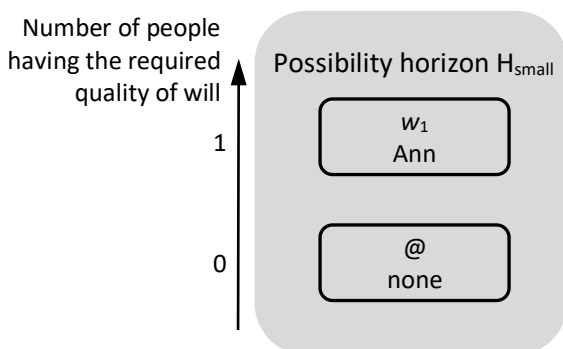
Now consider Ann. When we use  $H_{large}$ , we find that Ann is blameworthy for the collapse of the ecosystem. (i) The collapse of the ecosystem *just is* worse than its

<sup>25</sup> This depends on which account of intentional collective action we use. However, according to the standard accounts of intentional collective action (Searle 1990; Bratman 2014; Gilbert 2014), they did not perform a collective action.

<sup>26</sup> See French (1984); List and Pettit (2011).

continuing to thrive, and (ii) Ann’s poor quality of will at  $t$  in relation to the ecosystem’s collapsing versus continuing to thrive caused the ecosystem to collapse. (a) Her poor quality of will at time  $t$  is process-connected to the collapse of the ecosystem: her poor quality of will is, together with for instance Beth’s poor quality of will, minimally sufficient for the collapse of the ecosystem (likewise, Ann’s and Claire’s poor qualities of will are minimally sufficient for the collapse of the ecosystem), and this connection remains when we consider more intermediate times. (b) The collapse of the ecosystem is *less secure* at  $t$  in the closest world where Ann does not have a poor quality of will, i.e. in world  $w_1$ , where she has the required quality of will. Here, only one instead of both of the other boat owners would also need to have the required quality of will in order for the ecosystem not to collapse. For similar reasons, the ecosystem’s continuing to thrive is *more secure* at  $t$  in the closest worlds where Ann has the required quality of will. Thus, BLAMEWORTHINESS FOR yields the result that Ann is blameworthy for the ecosystem’s collapsing rather than continuing to thrive. The same applies to Beth and Claire.

However, Ann might argue that  $H_{\text{large}}$  is not the relevant possibility horizon. As we saw above, she might defend herself by saying something like the following: “given that Beth and Claire didn’t care...”. There is a straightforward reading of this statement in terms of which worlds should be included in the relevant possibility horizon: Ann is saying, essentially, that we should hold fixed that the others did not care enough about life in the lake to use the environmentally friendly but more expensive paint, and treat this as a background condition. If we comply with this request, we get a much smaller possibility horizon,  $H_{\text{small}}$ . This possibility horizon contains just two possible worlds: the actual world, and a world where Ann has the minimally required quality of will. The ecosystem collapses in both.



The collapse of the ecosystem is *just as secure* at  $t$  in the actual world as it is in the closest world(s) in  $H_{\text{small}}$  where Ann has the required quality of will, namely infinitely secure. Thus, condition (b) of BLAMEWORTHINESS FOR is not satisfied.

When we consider  $H_{\text{small}}$ , which holds fixed the motivations of the other boat owners, we thus find that Ann is *not* blameworthy for the collapse of the ecosystem.

This suggests that our conflicting intuitions about THE LAKE can be explained as the result of employing two different possibility horizons: employing the large possibility horizon  $H_{\text{large}}$  yields the result that Ann (as well as Beth and Claire) is individually blameworthy for the collapse of the ecosystem. Employing the small possibility horizon  $H_{\text{small}}$ , which treats Ann's quality of will as variable while holding the quality of will of the other boat owners fixed, yields the result that Ann is *not* individually blameworthy for the collapse of the ecosystem. Thus, it matters to our assessment of Ann's *individual* responsibility whether we consider a possibility horizon that treats her as part of a group where each member's quality of will is a candidate cause, or a possibility horizon that treats her poor quality of will as the *only* candidate cause. The conflicting intuitions are explained, not as arising from two kinds of entities that can be blameworthy – the group and the individual – but instead as arising from two different perspectives we can take when we are assessing the blameworthiness of an individual.

Of course, we cannot say that Ann both is and is not individually blameworthy for the collapse of the ecosystem. We need to single out one of the two candidate possibility horizons as being the one that is relevant to assessing Ann's blameworthiness. We have already suggested in RELEVANT POSSIBILITIES FOR BLAME that we should treat it as a relevant possibility that everyone involved could have had the required quality of will. If this is accurate, the larger possibility horizon is the correct one. We will now support this with two considerations.

First, given that it is important to decide what brought about the collapse of the ecosystem, we have reasons to widen and adjust  $H_{\text{small}}$ , while  $H_{\text{large}}$  can do the job satisfactorily. If we address Ann alone, her defence that the ecosystem would have collapsed whether or not she had cared as she should seems persuasive. At least, it seems to be an open question whether we should consider the smaller possibility horizon she insists on, or the larger one. However, Beth could also argue in the same way that *she* is not blameworthy for the collapse of the ecosystem. And so could Claire. The situation is symmetrical, so if the defence is open to one, it is open to each. If we accept Ann's defence and follow this argument to its logical conclusion, we end up concluding that neither Ann, nor Beth, nor Claire is a cause of the collapse of the ecosystem, and therefore none of them is blameworthy. At this point, we might suspect that something has gone wrong. Surely, we might think, Ann, Beth or Claire (or all three) must be blameworthy for the collapse of the ecosystem. After all, as Björnsson (2011) points out, it seems that *their* lack of care for the ecosystem played a crucial role in bringing about its collapse. So, we face two (related) puzzles: one concerning causation and one concerning blameworthiness. When facing such puzzles, we think there are reasons to reconsider one's choice of possibility horizon:

WIDENING AND ADJUSTING: If it is important to decide what brought about X, and if the causes of X are unsatisfactorily explained, we have reasons to look for more possibilities to include in the possibility horizon under consideration, and to scrutinise the possibilities we already have included.

So, once we realise that each of Ann, Beth and Claire could appeal to the DEFENCE – once we realise that the collapse of the ecosystem is unsatisfactorily explained – we will also realise that we have reason to broaden our possibility horizon (given that it is important to decide what brought about the collapse of the ecosystem). That is, we have reason to turn to the larger possibility horizon  $H_{\text{large}}$ , according to which all three are causes of and blameworthy for the collapse of the ecosystem.

Importantly, the reason to widen and adjust our possibility horizon is not grounded in the fact that, if we were to widen the possibility horizon, we could hold Ann, Beth and Claire responsible for the collapse, or say that they caused the collapse. If so, the reasoning would be circular. Rather, the reason to widen and adjust our possibility horizon is grounded in the fact that it is important to find out what brought about the collapse of the ecosystem. Maybe we seek to ensure that similar things do not happen in the future, or maybe we seek to find out who, if anyone, is to blame for the collapse of the ecosystem. If it would turn out upon closer scrutiny – even after widening the possibility horizon – that only Ann caused the collapse, or that none of them did (perhaps something else entirely caused the collapse), this is the conclusion we should accept.

Second, the smaller possibility horizons that Ann, Beth and Claire must appeal to in their defences do not fit together. As stated before, if we address Ann alone, her defence seems persuasive. However, this defence is only open to Ann as long as we treat Beth and Claire's quality of will as fixed. Ann's defence presupposes that *her* poor quality of will is a candidate cause, while Beth's and Claire's poor qualities of will are mere background conditions. Beth's defence, on the other hand, presupposes that *her* poor quality of will is a candidate cause, while the poor quality of will of Ann and Claire are background conditions. In order to accept Ann, Beth and Claire's defences, we have to adopt one perspective when evaluating whether Ann is blameworthy, a different perspective when evaluating whether Beth is blameworthy, and a third when evaluating Claire's blameworthiness. But these perspectives do not fit together. We cannot at the same time consider, for instance, Ann's quality of will as both a background condition and as a candidate cause.<sup>27</sup> Given the social function of blame, it should be possible to assess blameworthiness from a single perspective that everyone involved in the situation can accept, and that can be used to assess the blameworthiness of everyone involved. This disqualifies  $H_{\text{small}}$  since it does not provide such a perspective.

---

<sup>27</sup> Jamieson (2007: 176) proposes a similar argument.

We should point out that there would be no problem of possibility horizons not fitting together if we were to consider a possibility horizon where the poor quality of will of *each* of Ann, Beth, and Claire was treated as a mere background condition. If we for instance were to evaluate whether the company producing the hazardous paint is blameworthy for the collapse of the ecosystem, we might for the sake of simplicity treat Ann, Beth and Claire (and their qualities of will) as mere background conditions and instead focus on the quality of will of the company. Thus, there are in fact two ways to treat Ann, Beth, and Claire equally: treating them all as potential causes (as in  $H_{\text{large}}$ ), or treating them all as mere background conditions. However, when the question at issue is whether for instance Ann is to blame for the collapse of the ecosystem, there is a compelling reason not to treat Ann's poor quality of will as a mere background condition: doing so would *pre-judge* the question of whether she is blameworthy, by not treating her poor quality of will as a candidate cause at all. Furthermore, we would be failing to treat Ann as an agent by rejecting the possibility that she could have had the required quality of will. Thus, when Ann's blameworthiness is at issue, we have to treat Ann's poor quality of will as a potential cause; and then we also have to treat Beth and Claire's quality of will as potential causes. The relevant choice, then, is between  $H_{\text{small}}$  and  $H_{\text{large}}$ .

We therefore conclude that  $H_{\text{large}}$  is the relevant possibility horizon. Thus, BLAMEWORTHINESS FOR entails that Ann is directly individually blameworthy for the collapse of the ecosystem. By contrast, writers like Held (1970) and Wringer (2016) would argue that Ann is only *indirectly* individually blameworthy for this outcome. On their view, individual blameworthiness in cases like THE LAKE derives from the blameworthiness of the group. This presupposes that also unstructured groups – that is, groups that lack collective intentions or formal decision procedures – can be blameworthy. BLAMEWORTHINESS FOR does not presuppose this controversial idea. Still, our view captures Held's and Wringer's crucial insight that individual blameworthiness disappears from the scene if we lose sight of the group. For instance, if we fail to consider the possibility that each of Ann, Beth and Claire could have cared enough about life in the lake when we assess the blameworthiness of Ann, we will end up with  $H_{\text{small}}$  according to which Ann is not blameworthy for the collapse of the ecosystem.

## 7. Conclusion

We have suggested that BLAMEWORTHINESS FOR together with CAUSATION gives an accurate account of when you are blameworthy for an action, omission or outcome. This account captures the intuitive idea that you are blameworthy for something if this thing happened because you did not care enough.

One virtue of the account is that it gives the right verdict in a wide range of cases, including cases of forgetting, making a negative difference to a nevertheless good result (as when Sally did not pay attention to the chilli recipe), or doing an action with runaway consequences (like TRAGEDY), cases where we disregard irrelevant possibilities (like when we think the queen of Sweden did not cause Selma's flowers to die), omission cases (like INDIFFERENT JOHN), switching cases (like TROLLEY TROUBLE), (early) pre-emption cases (like BACKUP BILLY), Frankfurt-style cases (like BACKUP NEUROSCIENTIST BILLY) and collective harm cases (like THE LAKE). In addition, it also gives the right verdict in a wide range of other cases, including cases of overdetermination, late pre-emption, double prevention, and collective harm cases without a threshold, which we do not have space to discuss here.

Another virtue of the account is that it can explain the conflicting intuitions we have about some cases. In collective harm cases like THE LAKE, for instance, it explains why it might be tempting to accept Ann's defence that we should not blame her since, given that the other two boat owners did not care, her lack of care did not make any difference. If we treat Beth's and Claire's poor qualities of will as mere background conditions, as Ann insists we should, Ann is correct that her lack of care did not cause the collapse. At the same time, our account also explains why it seems that Ann, Beth and Claire *are* individually blameworthy for the collapse: when we treat it as an open possibility that each could have had the required quality of will, we find that each of them caused the bad outcome.

As might be evident by now, our account entails that it matters for blameworthiness for actions (etc.) what possibilities are relevant. Correspondingly, it matters for our judgements about blameworthiness which possibilities we consider to be relevant. This explains, for example, why our intuitions are torn in THE LAKE: we are torn between only treating it as relevant that Ann could have had the required quality of will, and treating it as relevant that each of the three could have had the required quality of will. To say something about blameworthiness itself, rather than merely about our judgements, we need to answer the question: which possibilities are relevant? We have argued that the relevant possibility horizon includes, as a minimum, every combination of the actual poor quality of will of the agents involved in the situation, and the quality of will they were minimally required to have. This means, for example, that the relevant possibility horizon in THE LAKE is the larger one, and so Ann, Beth, and Claire are blameworthy for the collapse.



## 12. Elaborating BLAMEWORTHINESS FOR

In this chapter, I elaborate and elucidate some aspects of BLAMEWORTHINESS FOR. First, as it stands, BLAMEWORTHINESS FOR gives counterintuitive verdicts in cases of deviant causation. To avoid this, we need to say that you are only blameworthy for X if your poor quality of will towards X was a *non-deviant* cause of X. Second, we used a simplified understanding of process-connections in the previous chapter. This understanding gives the wrong verdict in late pre-emption cases. Here, I set out an account of process-connections that avoids this failing. Third, process-connections and NESS seem quite similar. Both are accounts of minimal sufficiency. One might therefore wonder why we appeal to process-connections in CAUSATION instead of the better-known NESS condition. I argue that NESS either faces counterexamples or is underexplained, depending on how you spell out the finer details of the account. Fourth, in “You Just Didn’t Care Enough”, we used a simplified version of CAUSATION that is contrastive only on the effect side, not the cause side (“C is a cause of E rather than E\* ...”). Here, I supplement CAUSATION so that it is contrastive on both the effect side and cause side, and I update BLAMEWORTHINESS FOR to match this improvement. Finally, I discuss the in-virtue-of-relation in condition (i) of BLAMEWORTHINESS FOR. I argue, roughly, that X is bad in virtue of Y just in case X increases the security of Y and Y is bad. For instance, in SHOOTING AND POISONING, C’s poisoning P’s tea is bad in virtue of P’s death being bad because C’s poisoning P’s tea increases the security of P’s death. BLAMEWORTHINESS FOR then says that C is blameworthy for poisoning P’s tea since his doing so increased the security of the death of P (which is a bad outcome), and since his poor quality of will towards P was a cause of his poisoning the tea. With this in place, I show that (i) is equivalent to stating that you are blameworthy for X only if you had a Y-related reason not to X. In SHOOTING AND POISONING, (i) amounts to the claim that C is blameworthy for poisoning P’s tea only if he had a death-of-P-related reason not to poison the tea. BLAMEWORTHINESS FOR then says that C is blameworthy for poisoning P’s tea because he had a death-of-P-related reason not to poison the tea, and because his poor quality of will towards P was a cause of his poisoning the tea.



## Deviant Causation

As we noticed in “You Just Didn’t Care Enough”,<sup>1</sup> it is necessary to modify BLAMEWORTHINESS FOR to handle cases of deviant causation (sometimes called “wayward causation”). In many cases of deviant causation a planned outcome is brought about in an unplanned manner, as in the following example suggested by Sara Bernstein (2019):

[ANGRY CASSOWARY:] Suppose that Assassin shoots at Victim, intending to kill him, but the shot misses. However, the shot startles a sleeping cassowary who then angrily mauls Victim to death.

(Bernstein 2019: 151)<sup>2</sup>

Cassowaries are flightless birds that are very wary of humans. If provoked, they can inflict serious injuries, even fatal ones.

Assuming Assassin had no way of foreseeing that the cassowary, if startled, would maul the Victim to death, it seems wrong to blame him for Victim’s death. As Bernstein argues, he might be blameworthy for something else, such as intending to kill Victim or trying to do so, but surely he cannot be blamed for the killing. However, unless BLAMEWORTHINESS FOR is amended in some way or another, it entails that Assassin is blameworthy for Victim’s dying rather than surviving. Assassin’s poor quality of will is a cause of Victim’s dying rather than surviving, and we assume that Victim’s dying just is worse than his surviving.

So, we need to modify BLAMEWORTHINESS FOR. This can readily be done by requiring, roughly, that you are blameworthy for a bad outcome if and only if your poor quality of will in relation to this outcome *non-deviantly* caused it. More precisely, BLAMEWORTHINESS FOR should be modified to say the following:

BLAMEWORTHINESS FOR: you are blameworthy for X rather than X\* just in case there is a Y and Y\*, such that

- (i) X is worse than X\*, at least partly in virtue of Y being worse than Y\*, and
- (ii) there is a time *t* such that your poor quality of will at *t* in relation to Y versus Y\* is a *non-deviant* cause of X rather than X\* within the relevant possibility horizon H.

---

<sup>1</sup> In footnote 18.

<sup>2</sup> Similar examples were earlier proposed by Bennett (1965) and Davidson (1963).

This modification avoids the problem posed by deviant causation, but it is still not fully satisfactory. Unless we have a principled way of distinguishing deviant causation from non-deviant causation, we cannot with certainty say when condition (ii) of BLAMEWORTHINESS FOR is satisfied. There is currently no consensus on how such a distinction is to be made. Numerous proposals have been made, but these face numerous problems. Still, the debate on deviant causation might give us some direction.

Cases of deviant causation are often presented as counterexamples to causal theories of intentional action. Typically, causal theories state that an action is intentional if and only if it is caused by an intention, or by a combination of beliefs and desires, as in Donald Davidson's (1963) analysis.<sup>3</sup> Cases of deviant causation show that an action caused by an intention need not be intentional. For instance, while causal theories of intentional action entail that Assassin intentionally killed Victim in ANGRY CASSOWARY, intuitively he did not.

Some proponents of causal theories of intentional action, such as Myles Brand (1984) and Alfred R. Mele (1992), have argued that the relevant intention must be the proximate cause of action. This might do the trick in ANGRY CASSOWARY. In this case, we can see that Assassin's intention is not the proximate cause of the death of Victim; the cassowary's mauling is. Therefore, it is inferred, Assassin does not kill Victim intentionally.

Still, finding a satisfying account of proximate causation might be hard. For example, even in a simple assassination case without deviant causation (where the assassin shoots the victim and the shot is lethal) one might wonder whether the assassin's intention is a proximate cause, or whether instead the proximate cause is the presence of the flying bullet near the victim's body. If the latter, we will conclude that the assassin's intention is not the proximate cause of the victim's death, and hence that the assassin did not kill the victim intentionally. This is surely wrong.<sup>4</sup>

Moreover, the proximate causation strategy does not go well together with BLAMEWORTHINESS FOR. In some cases, we might think that someone is blameworthy for some outcome even if the causal pathway goes through the agency of others. This might be the case if someone hires an assassin to kill her enemy. Say that Dr. Evil hires an assassin to kill Mr. Good, and that the assassin succeeds in this endeavour. In that case, it seems that Dr. Evil is blameworthy for the death of Mr. Good. However, this is not the verdict we get if we say that you are blameworthy

---

<sup>3</sup> This problem was first proposed by Chisholm (1964a). For a clearer depiction of causal theories of intentional action than I have presented here, see Davidson (1963). For an extended defence of the idea that actions caused by intentions are intentional actions, see Thalberg (1984).

<sup>4</sup> Moreover, it is far from clear that the requirement that we must only consider proximate causes helps proponents of causal theories of action to avoid the problem posed by cases of deviant causation. See Mayr (2011). Mele later revised his views on this issue (see Mele 2003).

for an outcome only if your intention was the proximate cause of this outcome. At best, we can say that the assassin's intentions are the proximate cause of Mr. Good's premature death.

Perhaps causation is deviant when the planned outcome *occurs in an unplanned manner*, as Adam Morton (1975) and John Bishop (1989) have suggested.<sup>5</sup> For instance, about ANGRY CASSOWARY one might suggest that Assassin is not blameworthy for Victim's death, because Victim did not die in the way Assassin intended. Victim died by being mauled to death by the bird rather than by having his body pierced by the bullet. Although this proposal gives the right verdict in some cases, it fails to do that in others. As Bernstein (2019) argues, it gives the wrong verdict in cases where the deviantly produced outcome occurs in the same manner as the agent had planned, as happens in the following case:

LAZY ASSASSIN: Becky is dispatched to kill Victim. Becky shoots at Victim, but misses. The shot wakes up Bill, who was independently dispatched by a different assassination agency. Bill shoots Victim and kills him. (Bill would not have shot had Becky not awakened him.) Had Becky not shot, Victim would not have died.

(Bernstein 2019: 157)

Here, Victim dies by gunshot, just as Becky intended. Still, it seems that Becky is not blameworthy for killing Victim. Bill is.<sup>6</sup>

One might instead suggest that Assassin is not blameworthy for Victim's death in ANGRY CASSOWARY because the causal route from Assassin's pulling the trigger of his gun deviates from the one that he envisaged when pulling the trigger (and that the same could be said about Becky's case). The planned outcome is *brought about in an unplanned manner*, so to speak. As Bernstein (2019) puts it: "If an agent can't foresee the bizarre causal process leading to the outcome, then she is not responsible for it, or is at least less responsible for it" (155).

While it is more promising than the previous strategy, this suggestion also faces problems. For one thing, we now need a principled way to distinguish insignificant deviations from significant ones. Consider the following case:

---

<sup>5</sup> Their proposals differ in detail. For our purposes, however, this rough description of their suggestions will suffice.

<sup>6</sup> The proponents of the strategy of matching plans to the manner in which the outcome occurs might have resources to answer to this objection. Thalberg (1984), for instance, also requires the causal chains (deviant or not) not to go through the agency of others. With this proviso, he could maintain that Becky did not intentionally kill Victim in LAZY ASSASSIN on the grounds that her intention did not cause Victim's death. The alleged causal connection runs through the agency of Bill. However, this proposal faces other problems, one of which is that it leaves those who hire an assassin blameless for the deaths their assassins bring about. See Mayr (2011: ch. 5) for a more detailed critique.

GUST OF WIND: Assassin shoots at Victim, intending to kill him, and succeeds. However, while Assassin envisaged the bullet flying straight towards Victim, it did not do so. A gust of wind blew it slightly off the envisaged track, and another gust of wind then blew it back on track before it reached Victim at the indented spot.

In this case, Assassin seems blameworthy for killing Victim even though the bullet's actual route deviated somewhat from the one that Assassin intended. So, in order to make this strategy work, we must say that, in ANGRY CASSOWARY, the actual causal route deviates enough from what Assassin intended to count as a deviant causal route, while in GUST OF WIND, it does not. To find a principled way to distinguish deviant causes from nondeviant ones, that is, we will need a principled way to distinguish significant from nonsignificant differences between the planned causal route and the one that actually ensues.<sup>7</sup> Hopefully, this will be elaborated in the future. In the meantime, we will have to manage with the following imprecise idea:

#1 NON-DEVIANT CAUSE: your intention non-deviantly causes the planned outcome iff the planned outcome is brought about roughly as you had planned.

So far, we have considered cases where the deviant causal connection is found between the action and the consequence; Assassin fires the gun, the bullet misses its target, but because of some bizarre coincidence, his firing of the gun causes the intended outcome anyway. Brand (1984) helpfully labels such cases *cases of consequential waywardness*.<sup>8</sup> In a rather different kind of case, the deviant causal connection is located between the intention and the resulting behaviour (whether action or omission). The following case is borrowed from Mele:

[TREMBLING CHEMIST: A] chemist who is working with cyanide near his colleague's cup of tea may desire to kill his colleague and believe that he can do this by dropping some cyanide into the tea.... [T]his desire and belief may so upset him that his hands shake, with the result that he drops some of the poisonous substance into the tea.

(Mele 1983: 346)<sup>9</sup>

Here, the chemist intends to poison the tea, and his intention causes the tea to be poisoned, but it does not seem that he intentionally poisoned the tea. The planned

---

<sup>7</sup> A similar objection could be launched against Morton's and Bishop's idea that the manner in which the outcome occurs must match the agent's intention. Consider a case where Assassin succeeds in killing Victim, but where, while he had planned to hit him in the right eye, he actually hits him in the left. This small deviation from the plan does not seem to exculpate Assassin.

<sup>8</sup> To be precise, Brand labels the problem that such cases give rise to "the problem of consequential waywardness". See Brand (1984).

<sup>9</sup> This kind of example was first proposed by Morton (1975).

outcome is brought about in the wrong way. The chemist did not plan to become so excited that he accidentally drops poison into the tea. Unlike in ANGRY CASSOWARY and LAZY ASSASSIN, however, causation does not go astray after the relevant action has been performed; it goes astray even before that action takes place. Brand labels such cases *cases of antecedential waywardness*. The earlier suggestion, that your intention non-deviantly causes an outcome iff this outcome was brought about roughly as planned, can easily be extended to cover these cases as well. We could say that:

#2 NON-DEVIANT CAUSE: your intention non-deviantly causes the planned action, omission or outcome iff the planned action, omission or outcome is brought about roughly as you had planned.

This understanding of non-deviant causation gives the right verdict in TREMBLING CHEMIST. Intuitively, in this case, the chemist is not blameworthy for poisoning his colleague's tea. The poisoning of the tea is brought about in too accidental a manner for this to be the case. Granted, the chemist is blameworthy for something – for instance, for intending to poison his colleague's tea. But he cannot be blamed for poisoning the tea. This is also the verdict that BLAMEWORTHINESS FOR gives once it is modified to exclude cases of deviant causation. While the chemist's poor quality of will in relation to his colleague caused the tea's being poisoned, it did so in a deviant way: the poisoning of the tea was not brought about in roughly the intended way.

However, not all cases of deviant causation are cases where the planned action, omission or outcome is brought about in an unplanned manner. In some cases, causation goes astray *before* any plan is formed. Mele (1987) labels this kind of waywardness *tertiary waywardness*. Consider the following case:

KILLING KYLE: Suzy hates Kyle for no good reason. Given some time, she would eventually decide to kill him. However, unbeknown to her, the neuroscientist Billy is monitoring her processes of deliberation. Billy is in love with Suzy, and upon seeing her attitude to Kyle he decides to kill Kyle on his own. Billy has no grudge against Kyle, and had Suzy shown no ill will towards Kyle, Billy would not have killed him.

Here, the intuitive verdict is that Suzy is not blameworthy for Kyle's premature death (Billy is). We might blame Suzy for hating Kyle, but this does not make her blameworthy for killing him. Almost certainly, we will want to say that the reason why Suzy is not blameworthy for Kyle's death is that there is something deviant about the way Kyle's death was brought about. But the current suggestion about how to distinguish deviant causes from nondeviant ones does not help us out here. It concerns causation that goes astray *after* the agent has formed an intention to perform the relevant action, or to bring about the outcome in question, and in

KILLING KYLE causation goes astray before that. In order to get a better grip on deviant causation, we shall have to say something about tertiary waywardness.

Generalising from KILLING KYLE, one might suggest that Suzy's poor quality of will caused Kyle's death in a deviant way because her agency was by-passed. Her poor quality of will towards Kyle did not bring about an intention to kill Kyle: she did not deliberate over whether to kill Kyle or not, she did not make any decision to kill Kyle, and she did not perform any action with the intention of killing Kyle. Still, there may be cases of tertiary waywardness that do not fit this bill.<sup>10</sup> Until we have a better, more principled way to separate non-deviant causes from deviant ones, we will have to manage with something like the following crude test:

NON-DEVIANT CAUSE: In cases where you have an intention to X (perform an action, omit doing something or to bring about an outcome), your quality of will is a non-deviant cause of X iff (i) X is brought about roughly as you had planned, and (ii) causation does not go astray before you form your intention.

In cases where you do not have an intention to X, your quality of will is a non-deviant cause of X iff causation does not go astray.

Note that on this (admittedly sketchy) definition, your poor quality of will might be a non-deviant cause of some outcome even if you do not end up forming any intention to bring it about. So, just as the original version of BLAMEWORTHINESS FOR did, the revised version entails that you can be blameworthy for an outcome when you do not form any intention to bring it about. Some examples: you might be blameworthy for forgetting your best friend's birthday, if you did so, because you did not care as required about it, as happens in Angela Smith's (2005) famous case, and you might be blameworthy for forgetting to buy milk on your way home, if you do so, because you did not care as required about buying milk (or about the person who asked you to buy milk), as happens in Randolph Clarke's (2014) also famous case.

To sum up, BLAMEWORTHINESS FOR must be amended so that it does not give the wrong verdict in cases of deviant causation. Unfortunately, it is hard to find a principled way to isolate the deviance in question. #2 NON-DEVIANT CAUSATION gives some guidance in cases of antecedential and consequential waywardness (TREMBLING CHEMIST and ANGRY CASSOWARY) but not in cases of tertiary waywardness (KILLING KYLE). In the latter, at least for the time being, we will have to rely on our intuitive sense of when a cause is deviant (or is not). These results, as it were, are captured in NON-DEVIANT CAUSE.

---

<sup>10</sup> For instance, it is unclear whether Mele's (1987) original case of tertiary waywardness fits this bill.

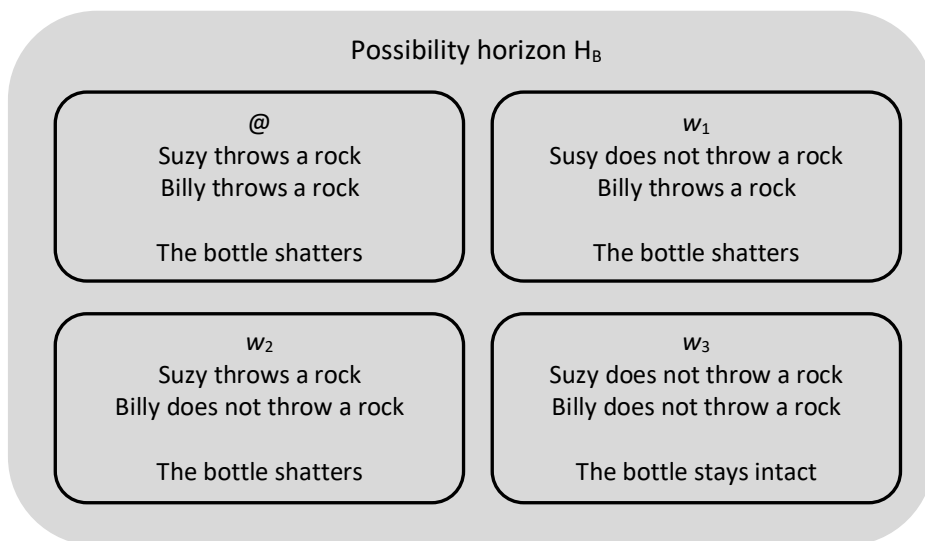
## Understanding Process-Connections

In “You Just Didn’t Care Enough”, we said that our account of process-connections needed further refinement in order to be able to handle all cases of late pre-emption.<sup>11</sup> To see why further refinement is needed consider again:

[BOTTLE SHATTERING:] Billy and Suzy throw rocks at a bottle. Suzy throws first, or maybe she throws harder. Her rock arrives first. The bottle shatters. When Billy’s rock gets to where the bottle used to be, there is nothing there but flying shards of glass. Without Suzy’s throw, the impact of Billy’s rock on the intact bottle would have been one of the final steps in the causal chain from Billy’s throw to the shattering of the bottle. But, thanks to Suzy’s preempting throw, that impact never happens.<sup>12</sup>

(Lewis 2000: 184)

We would say that Suzy’s throw caused the bottle shattering, while Billy’s did not. Given the guidance we gave in the previous chapter, however, CAUSATION<sup>13</sup> seems to entail that Billy’s throw *did* cause the bottle shattering. Consider the time  $t$  when Suzy throws her rock. At this time, there are four possibilities, as given by the following possibility horizon:



<sup>11</sup> In footnote 7.

<sup>12</sup> This case was introduced in Chapter 3, p. 81.

<sup>13</sup> You find this principle on p. 228.

(a) Billy's throwing his rock might seem to be process-connected to the bottle shattering. His throwing belongs to a set of events that is minimally sufficient for the bottle to shatter, and this remains true when we consider more and more intermediate times between his throwing the rock and the bottle shattering. There is no intermediate time where the relevant set is not minimally sufficient for the bottle to shatter. Add to this that (b) Billy's throw makes the bottle shattering more secure and its staying intact less secure: in the closest-to-@-at- $t$  world in  $H_B$  where Billy does not throw his rock ( $w_2$ ), the bottle shattering is less secure at  $t$ , and the bottle's staying intact is more secure at  $t$ , than they are in @. So, does CAUSATION erroneously entail that Billy's throw caused the bottle shattering?

It does not. We obtain that verdict when we operate with an underdeveloped understanding of minimal sufficiency. CAUSATION does not refer to minimal sufficiency of the kind depicted in the previous paragraph, but rather to *time-sensitive* minimal sufficiency. In "You Just Didn't Care Enough" we chose not to dwell on time sensitivity. We felt that was an unnecessary complication given that we were not considering late pre-emption cases.

To get a grip on what time-sensitive minimal sufficiency is, we can start by noting that Billy's rock arrives too late to be a cause of the shattering. This thought is the initial motivation for appealing to time sensitivity. However, this thought has to be elaborated. One way to do this would be to follow David Lewis' (2004) lead and consider whether Billy's throw belongs to a set of events that is minimally sufficient for a maximally temporally fragile version of the bottle shattering. However, as Lewis also acknowledges, this strategy soon runs into trouble: Billy's throw also makes a difference, albeit tiny, to the timing of the shattering, and without it the set would no longer be sufficient to bring about the maximally fragile version of the bottle shattering that actually occurred. As an alternative, Touborg (2018) suggests that we need to pay attention to the initial temporally robust version of the shattering, the maximally temporally fragile version of the shattering, *and all the intermediately more or less temporally fragile versions of the shattering*. More precisely, the idea is the following:

TIME-SENSITIVE MINIMAL SUFFICIENCY: A set of contemporaneous events  $S$  is time-sensitively sufficient for a later event  $E$  just in case  $S$  is minimally sufficient for  $E$ , and, for every temporally more fragile version  $E^+$  of  $E$ ,  $S$  (or some more fragile version of  $S$ ) is also minimally sufficient for  $E^+$ .

This suggestion might seem surprising at first. If neither looking at the initial temporally robust version, nor looking at the maximally temporally fragile version does the trick, why would looking at both of these versions as well as all the intermediate versions do so?



The answer is that while Billy's throw guarantees (in the circumstances and with the laws of nature) that the bottle will shatter, and while his throw is necessary for the bottle shattering to occur exactly at the time it did, his throw neither guarantees nor is necessary for the bottle to shatter roughly at the time it did. To be more precise, Billy's throwing the rock belongs to a set of concurrent events that is minimally sufficient for the bottle shattering to occur sometime (if we look at the robust version of the shattering). It also belongs to a set of events that is minimally sufficient for the bottle shattering to occur at *exactly* the time it did (if we look at the maximally fragile version of the shattering – remember Lewis' (2000) point). However, if we consider a slightly less temporally fragile version of the shattering – say, one that essentially occurs within  $\pm 0.1$  second of its actual occurrence – Billy's throw does not belong to a set of events that is minimally sufficient for the occurrence of this event. His throw is either not hard enough or not early enough. His rock will arrive at the place where the bottle used to be too late to guarantee that the bottle will shatter within this time interval. Moreover, his throw is not necessary for the set of events guaranteeing that the bottle shattering will occur within this interval. The gravitational force his rock exerts on the timing of the shattering is too tiny for this. In the circumstances and under the laws of nature, Suzy's rock guarantees on its own that the bottle will shatter within this interval. So, by requiring that for each version of the bottle shattering – from the temporally robust version, through the intermediate versions, to the maximally temporally fragile version – Billy's throw must belong to a set of events that is minimally sufficient for *that* version, we generate the conclusion that Billy's throw is not process-connected to the outcome.

In contrast, Suzy's throw *is* minimally sufficient for the bottle shattering whichever temporally more fragile version of this outcome we consider. Her throw belongs to a set of events that is minimally sufficient for the shattering to occur at all. It belongs to a set of events that is minimally sufficient for the shattering to occur in an interval of, say,  $\pm 0.1$  second of its actual occurrence. And it belongs to a set of events that is minimally sufficient for the shattering to occur in an infinitesimally small interval including the time at which the shattering actually occurred.

## Process-Connections and NESS

Process-connections and NESS seem to describe the same causal concept. Both are about sufficiency. There is a process-connection iff *C* is *minimally sufficient* for *E*, and the NESS condition says that *C* must be *necessary for the sufficiency* of a set that is sufficient for *E*. One might then wonder why we appeal to process-connections in CAUSATION instead of the more established NESS condition. Could we not use the following account of causation?

CAUSATION\*: Suppose that C occurs at  $t$  and E occurs later. Then C is a cause of E rather than E\* within possibility horizon H just in case

- (a) C satisfies the NESS condition of being a cause of E, and
- (b) there is at least one world in H where C does not occur, and in the closest-to-@-at- $t$  world(s) in H where C does not occur E is *less secure* at  $t$  and E\* is *more secure* at  $t$  than they are in @.

In many cases, we could. This account gives intuitively correct verdicts in a large range of cases. It does so, for instance, in switching cases, and it allows us to distinguish between causes and background-conditions. That is, it gives intuitively correct verdicts in many cases where NESS, taken on its own, does not (see the discussion in Chapter 3, p. 87ff). Consider, for instance, the switching case THE ENGINEER, where an engineer flips a switch so that a train travels down the right-hand track instead of the left. The tracks reconverge up ahead, so the train arrives at its destination at the same time and in the same way.<sup>14</sup> Here, it seems that the engineer's flipping the switch cannot be a cause of the train's arriving late. However, NESS (taken on its own) entails that it is. To recap, NESS states that:

NESS: A condition  $c$  was a cause of a consequence  $e$  if and only if it was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the occurrence of  $e$ .

(Wright 2013: 18)<sup>15</sup>

In THE ENGINEER, the engineer's flipping of the switch was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the train's arrival at the station, and thus a cause of the train's arrival at the station.

Now, if we also require that a cause must increase the security of its effect, we escape this counterintuitive verdict. The engineer's flipping of the switch does not increase the security of the train's arrival at the station. If she had not flipped the switch, the train would have travelled down the left-hand track instead, and the train's arrival at the station would have been just as secure.

Still, CAUSATION\* faces counterexamples. Quite what those are depends on how we understand the notion that a set of events is *sufficient* for an outcome.

---

<sup>14</sup> This case was introduced in Chapter 1, p. 27, and further discussed in Chapter 3, p. 90f, and in Chapter 6, p. 141f.

<sup>15</sup> According to Wright (2013), this is a clarification of the original version of NESS as formulated in Wright (1985). As we will soon see, Wright (2013) revises the formulation of NESS further.

## Lawful Sufficiency

On a common interpretation, a set of events is sufficient for an outcome if it *guarantees* that the outcome will occur.<sup>16</sup> Following Richard Fumerton and Ken Kress (2001), we can label this understanding of sufficiency *lawful sufficiency*. Lawful sufficiency entails that NESS and CAUSATION\* give counterintuitive verdicts in some late pre-emption cases. To see this consider, first, BOTTLE SHATTERING, where Suzy and Billy throw rocks at a bottle, and Suzy's rock hits the bottle, shattering it, a moment before Billy's rock flies into where the bottle used to be. If "being sufficient" means "guarantees", it might seem that NESS entails the mistaken verdict that Billy's throw caused the bottle shattering. His throw seems to be a necessary element of a set that *guarantees*, given the laws of nature, that the bottle will shatter. Given the current wind conditions, the accuracy of this throw, the laws of nature, and so on, his throw guarantees that the bottle will shatter, and without his throw, the relevant set no longer guarantees this outcome. Further, it might seem that CAUSATION\* entails the same verdict. For it seems that Billy's throw satisfies the NESS condition of causation, and, as I argued earlier in this chapter, his throw makes the bottle shattering more secure.

However, the appearance that NESS entails that Billy's throw caused the bottle shattering is misleading. If we pay closer attention to the relevant set, we will find some elements that are not instantiated. Importantly, since Suzy throws her rock, the bottle will no longer be in place when Billy's rock flies in. By throwing her rock, Suzy prevents Billy's rock from reaching the bottle. So, the set for which Billy's throw is necessary only guarantees the bottle shattering if Suzy does not throw her rock. Now, since NESS says that a cause must be part of a set of *existing* antecedent events that guarantees the occurrence of the outcome, and since the only set of which Billy's throw was a necessary element contains at least one non-occurring event (Suzy's not throwing her rock), NESS does not entail that Billy's throw was a cause of the bottle shattering. For the same reason, CAUSATION\* does not entail that Billy's throw was a cause of the bottle shattering. Roughly, we can say that Billy's throw did not satisfy the NESS condition of being a cause of the bottle shattering because the causal process initiated by his throw was cut short before it ran to completion. Call this strategy of explaining (away) causation in late pre-emption cases *the strategy of appealing to missing events*.

L. A. Paul and Edward Hall (2013) show that this strategy is unsuccessful in some cases of late pre-emption. It works in cases where the pre-empting cause (e.g. Suzy's throw) prevents some element in the pre-empted set from being instantiated (e.g. the bottle's being there).<sup>17</sup> Differently put, the strategy works in cases where the

---

<sup>16</sup> See for instance Fumerton and Kress (2001), Thomson (2008) and M. S. Moore (2009: 474).

<sup>17</sup> See also Paul (1998). Paul and Hall (2013) makes this point in connection with, not NESS, but Ramachandran's (1997) M-set analysis of causation.

process initiated by the pre-empted potential cause is cut short before it runs to completion. However, not all pre-emption cases are like this. Consider for instance the following case:

[THE DING DING CASE:] Billy and Suzy throw rocks, this time not at a bottle, but at a bell. The bell rings twice in rapid succession: the first as a result of Suzy's throw, the second as a result of Billy's.

(Paul & Hall 2013: 101)

Here, problematically, Billy's throw satisfies the NESS condition where the *first* ringing is concerned. The strategy of appealing to missing events does not work since Suzy's throw does not prevent the bell from being there when Billy's rock arrives. Billy's throw is necessary for the sufficiency of a set of antecedent conditions that guarantees the first ringing. To see this more clearly, suppose the first ringing occurs at noon sharp and the second ringing occurs one second later. Suppose also that the ringings are somewhat robust events. That is, they would have been the same events even if they had occurred at a slightly different time or in a slightly different manner. The first ringing could have happened, say, any time between noon and 1 second past noon without being a different event. Similarly, the second ringing could have happened at any time between noon and 1 second past noon without being a different event. With this stipulation, Billy's throw (in the circumstances and with the laws of nature) guarantees the first ringing. The first ringing is, essentially, just a ringing that occurs between 12 noon and 1 second past 12 noon, and Billy's throw guarantees a ringing within that time-interval. So, strange as it might seem, NESS entails that Billy's throw was a cause of the first ringing. So does CAUSATION\*. *Ex hypothesi*, this is not the case.

However, CAUSATION gives the right verdict in this case. There is no process-connection between Billy's throw and the first ringing. While Billy's throw is minimally sufficient for the first ringing to occur sometime, it is not minimally sufficient for it to occur exactly at the time it did, i.e. at noon. That is, Billy's throw is not time-sensitively minimally sufficient for the first ringing. Therefore, process-connections are more accurate in pinpointing the relevant causal connection between cause and effect than the NESS condition is – at least, when we take the relevant notion of sufficiency to be lawful sufficiency.

## Causal Sufficiency

Richard Wright (1985, 1988, 2013) can avoid the difficulty created by late pre-emption in THE DING DING CASE because he has an alternative understanding of what it is for a set to be sufficient for an outcome. He argues that the relevant notion of

sufficiency is *causal sufficiency*.<sup>18</sup> There must be a *causal law* connecting a cause to its effect. The idea is roughly this: through experience and experiments we know that when certain conditions are satisfied this or that effect will immediately follow. For illustration, we know that a bottle will shatter if someone throws a rock towards it, the throw is accurate enough, the rock reaches the bottle, the bottle is made of glass, and so on. These are the conditions of the relevant causal law. More accurately, Wright defines causal law in the following way:

A causal law is an empirically derived statement that describes a successional relation between a set of abstract conditions (properties or features of possible events and states of affairs in our real world) that constitute the antecedent and one or more specified conditions of a distinct abstract event or state of affairs that constitute the consequent such that, regardless of the state of any other conditions, the instantiation of all the conditions in the antecedent entails the immediate instantiation of the consequent, which would not be entailed if less than all of the conditions in the antecedent were instantiated.

(Wright 2013: 19)

In general, we are in no position to recount all conditions in the relevant causal law. Instead, we make do with *causal generalisations* including particularly salient conditions. In BOTTLE SHATTERING, the relevant causal generalisation would remind of the list given above.<sup>19</sup> If we understand sufficiency in this way, we can explain why Billy's throw did not cause the first ding in THE DING DING CASE. One condition was not satisfied. Billy's rock never reached the bell. For the same reason, if we take premise (i) of CAUSATION\* to refer to Wright's version of the NESS condition, it will not follow that Billy's throw was a cause of the first ringing. While Billy's throw increases the security of the first ringing, it does not satisfy the NESS condition.

While Wright can avoid the problem illustrated by THE DING DING CASE, he runs into a different one. Consider BOTTLE SHATTERING once more. Here, we have assumed that the relevant causal law includes conditions like "someone throws a rock towards the bottle", "the throw is accurate enough", "the rock reaches the bottle", "the bottle is made of glass", and so on. However, upon closer scrutiny, Wright's definition of a causal law does not help us pin down all these conditions. Strictly speaking, the fact that someone is throwing the rock is not necessary for the sufficiency of any bottle shattering. It is not true that the instantiation of the bottle shattering "would not be entailed if less than all of the conditions in the antecedent were instantiated". It is enough that the rock comes flying, reaches the bottle (and

---

<sup>18</sup> This is particularly clear in Wright (2013).

<sup>19</sup> Wright's (2013) discussion of the case where two bridges have collapsed into a river and a ship is delayed as a result clearly illustrates this idea.

so on). The satisfaction of the conditions in this more restricted set entails that the instantiation of the bottle shattering will follow immediately. The reason why the rock comes flying is irrelevant.

Upon reflection, we realise that according to Wright's definition of a causal law, we need only consider the time slice immediately before the outcome occurs, and the conditions that were satisfied then. Any event that occurred earlier is redundant as regards the sufficiency of the relevant set. Problematically, if we do not include a condition like "someone throws a rock towards the bottle" in the relevant causal law, we lose our explanation of why Suzy's throw caused the bottle to shatter. It will then follow that her throw did not satisfy any condition in the relevant causal law, and so is not a cause of the bottle shattering.

Now, it is obvious that earlier events could be causes. Clearly, Suzy's throw was a cause of the bottle shattering. As clearly, Wright would take Suzy's throw to be part of the relevant causal generalisation. The problem is to explain why NESS, on the interpretation at hand, entails that we should include such conditions. There is nothing in the NESS formula – nor in Wright's definition of a causal law – that requires us to include earlier events like Suzy's throw in the relevant causal generalisation.

Perhaps there is a principled way to extend Wright's version of NESS so that it accurately identifies the conditions that should be included in the relevant causal law. If there is, process-connections and this version of NESS will be extensionally on a par (at least, as far as I can tell). At any rate, they will give the same verdicts in the cases we have considered here. That would allow us to substitute NESS for process-connections in CAUSATION. For now, however, it seems to me that process-connections capture the relevant causal relations more satisfyingly than NESS does.

## Causal Contrasts

You might have noticed that the security conditions in REASON and CAUSATION differ. The security conditions read as follows:

From CAUSATION:

[...] C is a cause of E rather than E\* within possibility horizon H just in case [...]

- (b) there is at least one world in H where C does not occur, and E is *less secure* at *t* and E\* is *more secure* at *t* in the closest-to-@-at-*t* world(s) in H where C does not occur than they are in @.<sup>20</sup>

---

<sup>20</sup> I have changed the order of some of the clauses to facilitate comparison.

From REASON:

You have a teleological reason to  $\varphi$  rather than  $\psi$  at time  $t$ , [...] just in case [...] there are two incompatible outcomes  $O$  and  $O^*$ , such that

- (d)  $O$  is *more secure* at  $t$  and  $O^*$  is *less secure* at  $t$  in the closest-to-@-at- $t$  world(s) in  $H$  where you  $\varphi$  at  $t$  than they are in the closest-to-@-at- $t$  world(s) in  $H$  where you  $\psi$  at  $t$  [...]

There are three differences I wish to highlight. First, the security condition of CAUSATION starts out requiring that “there is at least one world in  $H$  where  $C$  does not occur”. There is no corresponding requirement in the security condition of REASON. Second, the security condition of CAUSATION specifies that the relevant contrast on the cause side is the closest-to-@-at- $t$  world(s) in  $H$  where  $C$  does not occur, but leaves it up to us to decide which effects to compare (it leaves it up to us to decide  $E$  and  $E^*$ ). REASON, on the other hand, leaves it up to us to decide also what the relevant contrast is on the cause side. It takes as input both the relevant actions  $\varphi$  and  $\psi$ , and the relevant outcomes  $O$  and  $O^*$ . In short, we might say that while CAUSATION only is contrastive on the effect side, REASON is contrastive also on the cause side. Third, while the security condition of CAUSATION requires  $E$  to be *less secure* at  $t$  and  $E^*$  to be *more secure* at  $t$ , the security condition of REASON requires  $O$  to be *more secure*, and  $O^*$  *less secure*, at  $t$ .

These differences do not matter much, but they make comparisons between REASON and BLAMEWORTHINESS FOR (which builds on CAUSATION) difficult. In the next section, I will compare these two principles. To make this comparison easier, I will now streamline the formulations of the security condition.

First, CAUSATION’s requirement that there is at least one world in  $H$  where  $C$  does not occur is stated as the separate condition (a) in REASON, which says that there must be an option for you to  $\varphi$  at  $t$ . This amounts to saying that there must be at least one world in  $H$  where you  $\varphi$ . So, there is no substantive difference – only one of presentation. If condition (a) of REASON is satisfied, so is the first part of condition (b) of CAUSATION, and vice versa.

Second, it would have been more accurate to include contrasts also on the cause side of CAUSATION, just as we did with REASON. By not including a causal contrast  $C^*$  in CAUSATION in “You Just Didn’t Care Enough”, Touborg and I hoped to simplify things. When we are dealing with CAUSATION and BLAMEWORTHINESS FOR, explicit mention of the cause contrasts does not make any difference to the verdicts they deliver on blameworthiness. This is because we only include worlds in which the agent has his actual quality of will or the required quality of will in the relevant possibility horizon. Given this, there is only one alternative to the agent’s having his actual quality of will – namely, his having the required quality of will. Within the possibility horizons we are looking at, it therefore makes no difference whether

we consider the closest-to-@-at-*t* world(s) in H where C does not occur or the closest-to-@-at-*t* world(s) in H where C\* occurs. These are the same worlds. To put the same point differently, it does not matter whether we say “the agent’s having his actual quality of will *rather than not* was a cause of E rather than E\*” or “the agent’s having his actual quality of will *rather than the required quality of will* was a cause of E rather than E\*” In both cases we will be looking at the closest world(s) in H where the agent has the required quality of will.

To accommodate contrasts on the cause side of CAUSATION we have to make a few changes to condition (b). Besides requiring that there is at least one world in H where C does not occur, we must also require there to be at least one world in H where C\* does not occur. Moreover, instead of comparing the closest-to-@-at-*t* world(s) in H *where C does not occur* with @, we must compare the closest-to-@-at-*t* world(s) in H *where C\* occurs* with @. Finally, we must replace @ in the comparison with “the closest-to-@-at-*t* world(s) in which C occurs”. Since C is an event that occurs in the actual world, the closest-to-@-at-*t* world(s) in which C occurs is simply @, so this change does not make any difference to which causal verdicts CAUSATION gives.

Third, we now have a condition (b) of CAUSATION requiring that E is *less secure*, and E\* *more secure*, in the closest-to-@-at-*t* world(s) *where C\* occurs* than they are in the closest-to-@-at-*t* world(s) *where C occurs*. If we just turn things around, we will get a condition requiring that E is *more secure*, and E\* *less secure*, in the closest-to-@-at-*t* world(s) *where C occurs* than they are in the closest-to-@-at-*t* world(s) *where C\* occurs*. We have now finally arrived at a condition which closely resembles condition (d) of REASON, and which we can use when comparing the two principles. The updated version of CAUSATION reads as follows:

CAUSATION: Suppose that C occurs at *t* and E occurs later, that C\* is a merely possible event that is incompatible with C, and that E\* likewise is a merely possible event that is incompatible with E.

Then C rather than C\* is a cause of E rather than E\* within possibility horizon H just in case

- (a) C is process-connected to E,
- (b) there is at least one world in H where C occurs at *t*, and at least one world in H where C\* occurs at *t*, and E is more secure, and E\* is less secure, in the closest-to-@-at-*t* world(s) in H where C occurs at *t* than they are in the closest-to-@-at-*t* world(s) in H where C\* occurs at *t*.

When I refer to CAUSATION from now on, this is the version of the principle to which I am referring.

We should also revise BLAMEWORTHINESS FOR so that it is compatible with the updated version of CAUSATION. BLAMEWORTHINESS FOR must now specify that the



relevant contrast on the cause side is the possible world in which the agent has the required quality of will, as follows (with the changes italicised):

BLAMEWORTHINESS FOR: you are blameworthy for X rather than X\* just in case there is a Y and Y\*, such that

- (i) X is worse than X\* at least partly in virtue of Y being worse than Y\*, and
- (ii) there is a time *t*, such that your having a poor quality of will at *t* in relation to Y versus Y\* *rather than the required quality of will* is a non-deviant cause of X rather than X\*, within the relevant possibility horizon H.

## The in-Virtue-of Relation

Condition (i) of BLAMEWORTHINESS FOR requires that X is worse than X\* at least partly in virtue of Y being worse than Y\*. We might then ask: What does it mean that X is worse than X\* *partly in virtue of* Y being worse than Y\*? To get a grip on this question, consider the following late pre-emption version of BACKUP BILLY:

BELATED BILLY: Suzy is walking down the street. When she reaches the big house on the corner, she stops and considers. For no good reason, she has an intense dislike for the nice elderly couple who live in the house, and she has just got an idea: she is going to upset them by breaking their window on the first floor. Billy also dislikes the elderly couple, and has got the same idea independently of Suzy. Unbeknown to each other, each carefully selects a rock and hurls it towards the window. Suzy throws first, or maybe she throws harder. Her rock arrives first. The window shatters. When Billy's rock gets to where the window used to be, there is nothing there but flying shards of glass. Without Suzy's throw, the impact of Billy's rock on the intact window would have shattered the window. But, thanks to Suzy's pre-empting throw, that impact never happens.

Here, intuitively, both Suzy and Billy are blameworthy for throwing their rocks, but Suzy alone is blameworthy for breaking the window. Billy tried to break the window, but he failed. We might then ask in virtue of what Suzy and Billy are blameworthy for throwing their rocks. They do not seem blameworthy for doing so on the grounds that throwing rocks is *generally* worse than refraining to do so. To throw rocks might even be a nice thing to do, as it is when one skims stones at the beach on a beautiful summer's evening. Rather, they seem blameworthy for throwing their rocks because their doing so stands in some kind of relation to the elderly couple's becoming upset. This idea is captured in BLAMEWORTHINESS FOR, which states that they are blameworthy for throwing their rocks only if doing so is worse than not doing so at least *partly in virtue of* the elderly couple's becoming upset rather than not. But what is this "partly in virtue of" relation?

Let us start with the “partly”. The idea is that, on this occasion, Suzy’s throwing her rock is all-things-considered worse than her not throwing her rock, and one of the considerations pulling in this direction is that her throwing her rock rather than not doing so stands in the right relation to the elderly couple’s becoming upset rather than not doing so. The same goes for Billy. More generally, the idea is that Y’s being worse than Y\* is at least one of the considerations (maybe the only consideration) in virtue of which X is all-things-considered worse than X\*.

But what does it mean to say that X is worse than X\* at least partly “in virtue of” Y’s being worse than Y\*? Given the account of causation we are using there are two natural suggestions, each corresponding to a specification of condition (i) of BLAMEWORTHINESS FOR:

- X is worse than X\* since X rather than X\* *causes* Y rather than Y\*, and
- X is worse than X\* since X rather than X\* *makes it more secure* that Y rather than Y\* will occur.

One intuition supporting the first specification is that it seems right to say that an event is bad if it leads to a bad outcome. On the specification, one might say that X is instrumentally worse than X\*. If X rather than X\* occurs this will lead to Y rather than Y\* occurring, where Y is worse than Y\*.

An intuition supporting the second specification is that, roughly, it seems right to say that an event is bad if it risks leading to a bad outcome. On the specification, one might say that X is worse than X\* since if X rather than X\* occurs there is an increased risk that Y rather than Y\* will occur, where Y is worse than Y\*. Still, this depiction is very rough. I would prefer to talk about this in terms of increased security rather than in terms of increased risk.

So, which specification is the more accurate one? Let us decide by considering whether either of the specifications gives the right verdict on blameworthiness in BELATED BILLY.

BLAMEWORTHINESS FOR entails that Suzy is blameworthy for throwing her rock on both specifications of what it means to say that something is bad in virtue of something else’s being bad. (i) Her throwing the rock is worse than her not doing so, at least partly in virtue of the fact that the elderly couple’s becoming upset is worse than their not becoming upset. And this is true regardless of whether we think of her throwing of the rock as causing the elderly couple’s getting upset (which it does), or think of her throw as something that increases the security of the elderly couple’s becoming upset (which it also does: this follows from the claim about causation). Further, (ii) there is a time *t* such that Suzy’s having a poor quality of will towards the elderly couple rather than not having that quality of will is a non-deviant cause of her throwing the rock rather than not doing so.

It is less clear, however, that BLAMEWORTHINESS FOR shows Billy to be blameworthy for throwing his rock. Depending on our choice of specification of what it means for an outcome to be bad in virtue of another outcome, we get different verdicts. Suppose we understand (i) along the following lines: X is worse than X\* since X rather than X\* *causes* Y rather than Y\*. In that case, Billy's throwing his stone *is not worse* than his not throwing it, at least not in virtue of the elderly couple's becoming upset being worse than their not doing so. Billy's throw does not cause the elderly couple to be upset; Suzy's does.

Now imagine instead that we understand (i) along these lines: X is worse than X\* since X rather than X\* *makes it more secure* that Y rather than Y\* will occur. In that case, Billy's throwing *is worse* than this not throwing in virtue of the elderly couple's becoming upset being worse than their not doing so. His throwing his rock makes the elderly couple's becoming upset more secure. This suggests that we should understand (i) along these lines. This will give the right verdict in BELATED BILLY both when it comes to Suzy and when it comes to Billy. Thus, we can specify BLAMEWORTHINESS FOR in the following manner:

BLAMEWORTHINESS FOR: you are blameworthy for X rather than X\* just in case there is a Y and Y\*, such that

- (i) X is worse than X\* at least partly since
  - (i1) Y is worse than Y\*, and
  - (i2) there is at least one world in  $H_2$  where X occurs at  $t_2$ , and at least one world in  $H_2$  where X\* occurs at  $t_2$ , and Y is more secure and Y\* is less secure in the closest-to-@-at- $t_2$  world(s) in  $H_2$  where X occurs at  $t_2$  than they are in the closest-to-@-at- $t_2$  world(s) in  $H_2$  where X\* occurs at  $t_2$ .
- (ii) there is a time  $t$ , such that your having a poor quality of will at  $t$  in relation to Y versus Y\* rather the required quality of will is a non-deviant cause of X rather than X\* within the relevant possibility horizon H.

Here, (i2) is a more precisely formulated version of the idea that the occurrence of X rather than X\* makes it more secure that Y will occur rather than Y\*. Note that we have to use a slightly different possibility horizon (called  $H_2$ ) when we are deciding whether this is the case – different, that is from the one we use when deciding whether your poor quality of will is a non-deviant cause of the outcome. X and X\* do not occur at time  $t$ , when you could have had the required quality of will rather than your actual poor quality of will, but at some later time  $t_2$ .

Like the previous version of BLAMEWORTHINESS FOR, this one gives the right verdict across a wide range of cases. However, this version has the advantage of specifying more clearly what it is for an outcome to be good in virtue of some other outcome.

What should we say about the intuition supporting the first, now discarded specification – the intuition that it is right to say that an event is bad if it leads to a bad outcome? Actually, both specifications can explain this intuition (not only the first). On both, we arrive at the result that whenever X rather than X\* is a cause of Y rather than Y\*, and Y is worse than Y\*, then X is worse than X\* at least partly in virtue of Y being worse than Y\*. If X rather than X\* is a cause of Y rather than Y\*, then it is also true that X rather than X\* makes it more secure that Y rather than Y\* will occur, at least given CAUSATION.

### The Connection to REASON

Conditions (i1) and (i2) in this elaborated version of BLAMEWORTHINESS FOR correspond to the conditions laid down in REASON (see p. 125). Condition (i1) of BLAMEWORTHINESS FOR corresponds to condition (c) of REASON,<sup>21</sup> and condition (i2) of BLAMEWORTHINESS FOR corresponds to the three remaining conditions of REASON. So, we arrive at the proposition that X is worse than X\* in virtue of Y's being worse than Y\* just when there is an outcome-related reason not to X, but to X\* instead. In BELATED BILLY, for example, Billy is blameworthy for throwing his rock rather than refraining from doing so since he had an outcome-related reason – that is, a the-elderly-couple's-becoming-upset-related reason – not to throw the rock, but to refrain from doing so instead, and since his having a poor quality of will towards the elderly couple rather than not caused him to throw the rock rather than to refrain from doing so. More generally, the following version of BLAMEWORTHY FOR is analytically identical to the above one if REASON is assumed:

BLAMEWORTHINESS FOR: you are blameworthy for X rather than X\* just in case there is a Y and Y\*, such that

- (i) at the relevant time, you have a Y (vs Y\*)-related reason not to X, but to X\* instead.
- (ii) there is a time *t* such that your having a poor quality of will at *t* in relation to Y versus Y\* rather than the required quality of will is a non-deviant cause of X rather than X\* within the relevant possibility horizon H.

If we simplify things by disregarding the contrasts, we find that the following things are equivalent:

Billy is blameworthy for throwing the rock since his throwing of the rock was bad in virtue of the elderly couple's becoming upset, and since his poor quality of will towards the elderly couple was a non-deviant cause of his throwing the rock.

---

<sup>21</sup> Note, though, that it is necessary to set  $Y = O^*$  and  $Y^* = O$  so that Y's being worse than Y\* amounts to O's being better than O\*.

Billy is blameworthy for throwing the rock since his throwing the rock increased the security of the elderly couple's becoming upset, and since his poor quality of will towards the elderly couple was a non-deviant cause of his throwing the rock.

Billy is blameworthy for throwing the rock since he had a the-elderly-couple's-becoming-upset-related reason not to throw the rock, and since his poor quality of will towards the elderly couple was a non-deviant cause of his throwing the rock.

The final point nicely captures the intuition that you are only blameworthy for something if you had a reason not to do or cause this thing.

## A Potential Problem

Now that we have clarified what the in-virtue-of-relation consists in, we can see a potential counterexample to BLAMEWORTHINESS FOR. Consider again our switching case TROLLEY TROUBLE. In this, Suzy can flip a switch with the result that a trolley will travel down the left-hand track killing five people that are tied to the track. If she does not flip the switch, the trolley will travel down the right-hand track instead. Suzy wants the five to die for no good reason, and flips the switch. However, unbeknown to her, the left-hand track and the right-hand track converge before the track reaches the five. So, whether she flips the switch or not, the five will die.

Here, Suzy does not seem to be blameworthy for the five deaths. This is also the verdict BLAMEWORTHINESS FOR gives (as we argued in "You Just Didn't Care Enough"). We may well feel that she *is* blameworthy for *flipping the switch with the intention of killing the five*, but this verdict is not confirmed by BLAMEWORTHINESS FOR. Flipping the switch with the intention of killing the five does not increase the security of the five's being killed. Their death is just as secure whether Suzy flips the switch with the intention of killing the five or not.

I think we must separate two things here: the question whether Suzy is blameworthy for flipping the switch, and the question whether she is blameworthy for having the intention of killing the five. It seems that she is not blameworthy for the former but is blameworthy for the latter. This is the result BLAMEWORTHINESS FOR delivers. Flipping the switch does not increase the security of the five's being killed, so BLAMEWORTHINESS FOR entails that she is not blameworthy for flipping the switch.<sup>22</sup> However, BLAMEWORTHINESS FOR entails that she is blameworthy for having the intention to kill the five. Her having this intention is worse than her not having it in virtue of the five's being killed: if Suzy has the intention to kill the five, she will probably find some way or other of killing them. Intentions are not one-

---

<sup>22</sup> We should also separate a third issue. When she flips the switch with the intention of killing the five, Suzy reveals something about herself – that she is a bad person. And, she might be blameworthy in virtue of being a bad person even though she is not blameworthy for flipping the switch or for the death of the five. See the discussion in Chapter 10.

dimensional. For instance, if it turns out that the trolley derails before running over the five, Suzy might find some other way of killing them. In this way, her having this intention increases the security of the five's being killed – at least, if we widen the relevant possibility horizon in a plausible manner. This means that condition (i) of BLAMEWORTHINESS FOR is satisfied. Moreover, (ii) her poor quality of will towards the five was (we can assume) a non-deviant cause of her having this intention. Hence, both conditions of BLAMEWORTHINESS FOR are satisfied. So, BLAMEWORTHINESS FOR entails that she is blameworthy for having the intention to – in one way or another – kill the five.

I think these are the right verdicts: Suzy is not blameworthy for flipping the switch, but she is blameworthy for having the intention of killing the five. When we fail to separate these two things, as we do when we say Suzy is blameworthy for flipping the switch with the intention of killing the five, it is unclear exactly what our intuitions are tracking, and then BLAMEWORTHINESS FOR goes astray.

## Conclusion

In this chapter, I have elaborated and elucidated some aspects of CAUSATION and BLAMEWORTHINESS FOR. First, I gave a rough account of what it means for a person's quality of will to be the non-deviant cause of an outcome. Roughly, your poor quality of will non-deviantly causes some outcome if this outcome was brought about roughly as you had planned, and if causation did not go astray in some other way. Second, I went into the finer details of what it means for a cause to be process-connected to an outcome, and argued that process-connection better captures the idea that a cause must be connected to its effects than NESS does. Third, I modified CAUSATION and BLAMEWORTHINESS FOR to allow for causal contrasts on the cause side as well as the effect side in order to facilitate comparisons between these principles and REASON. Fourth, I suggested that an outcome is bad in virtue of some other outcome just in case it makes this outcome more secure, and I then showed that this understanding issues in some intuitively correct verdicts. Fifth, I showed that if we accept this understanding of what it is for an outcome to be bad in virtue of another outcome, it is possible to restate BLAMEWORTHINESS FOR to say, roughly, that you are blameworthy for X just in case you have a Y-related reason not to X, but your lack of care for Y makes you X anyway. Finally, I considered what might appear to be a counterexample to BLAMEWORTHINESS FOR. I argued that BLAMEWORTHINESS FOR is capable of explaining our intuitions in this case if we carefully distinguish what these intuitions track, or are about. In the next two chapters, which also are the final chapters of this thesis, I will put these clarifications and insights to work in the evaluation of a range of additional examples.



## 13. Applying BLAMEWORTHINESS FOR

One major virtue of BLAMEWORTHINESS FOR is that it gives intuitively correct verdicts about blameworthiness in a wide range of cases. In “You Just Didn’t Care Enough”, we showed that it gives intuitively correct verdicts in omission cases, switching cases, early pre-emption cases, Frankfurt-style cases, collective harm cases with a threshold, and in a number of other types of cases. In the previous chapter, I showed that it also gives the right verdict in late pre-emption cases. In this chapter, I will show that it gives the right verdicts in non-threshold cases, in the case of climate change, and in a case called “Penned-In Sharks” which provides difficulties for John Martin Fischer and Mark Ravizza’s (1998) influential account of moral responsibility. I will also consider a potential counterexample to the idea that a cause always increases the security of its effect.

### Non-threshold Cases

In “You Just Didn’t Care Enough”, we argued that BLAMEWORTHINESS FOR gives intuitively correct verdicts in collective harm cases with a threshold. We focused on THE LAKE, but the same argument applies to any collective harm case with a threshold, such as ASSASSINS and voting.

How about non-threshold cases like HARMLESS TORTURERS and DROPS OF WATER, where no act makes a perceptible difference to the outcome? According to some writers, there are no such cases. Alastair Norcross (1997) and Shelly Kagan (2011) argue that these are threshold cases in disguise. According to them, some flipping of a switch and some donation of a pint makes a perceptible difference in harm. Derek Parfit (1984), Zach Barnett (2018) and John Broome (2019) would agree that these are threshold cases in disguise, but they argue instead (in different ways) that each flipping of a switch and each donation makes a morally relevant difference in harm, although an imperceptible one.

If my arguments in “Making a Vague Difference” (Chapter 8) are correct, there might still be cases where no act makes a perceptible difference for the outcome. The arguments that Norcross and Kagan present to the contrary are mistaken. Barnett’s (2018) *no free lunch* argument, however, seems to establish that acting in the relevant way in collective harm cases makes a morally relevant difference, albeit



an imperceptible one (as argued in Chapter 9). Still, even if Barnett's argument seems to show that also imperceptible effects can be harms, it might eventually turn out that also this argument is flawed. In that case, we need an account that can deliver intuitively correct verdicts in non-threshold cases. And, as Parfit (1984: 82) argues, even if we believe that imperceptible differences might be harms, this is a controversial verdict, and some people disagree. Therefore, it is better if we have an account that can deliver the intuitively correct verdict in cases like HARMLESS TORTURERS and DROPS OF WATER without relying on the presumption that imperceptible difference might be harms.

BLAMEWORTHINESS FOR gives the intuitively correct verdict in non-threshold cases even if no act makes a perceptible difference in harm, and even if imperceptible effects cannot be morally relevant. It for instance gives the right verdict in DROPS OF WATER and HARMLESS TORTURERS even if no flipping of a switch or donation of a pint makes a perceptible difference in harm, and even if imperceptible harms are not morally relevant. In fact, BLAMEWORTHINESS FOR allows us to treat these cases in a way that closely parallels our treatment of THE LAKE. To make things easier, I will concentrate on one case. I will concentrate on HARMLESS TORTURERS (here presented in shortened form).<sup>1</sup>

HARMLESS TORTURERS. There are a thousand torturers and one victim. At the start of the day, the victim is already feeling mild pain. Each of the torturers flips a switch, making some instrument affect the victim's pain in a way that is imperceptible. After each torturer has flipped his switch, the victim is in excruciating pain.

Just as in THE LAKE, it seems that each torturer has a defence: "given that the other torturers did not care about the victim, my poor quality of will made no difference to the victim's pain". This defence asks us to consider a small possibility horizon where the poor quality of will of the other torturers is treated as mere background conditions. If we do, we will find that no torturer is blameworthy for the victim's being in excruciating pain. Given these small possibility horizons, it is not up in the air whether the victim will be in pain or not.<sup>2</sup>

As in THE LAKE, however, the correct possibility horizon treats each torturer's poor quality of will as a potential cause. For one thing, if each torturer defends himself in the mentioned way, and if we accept these defences, we must conclude that no torturer is blameworthy for the victim's being in excruciating pain, and that no

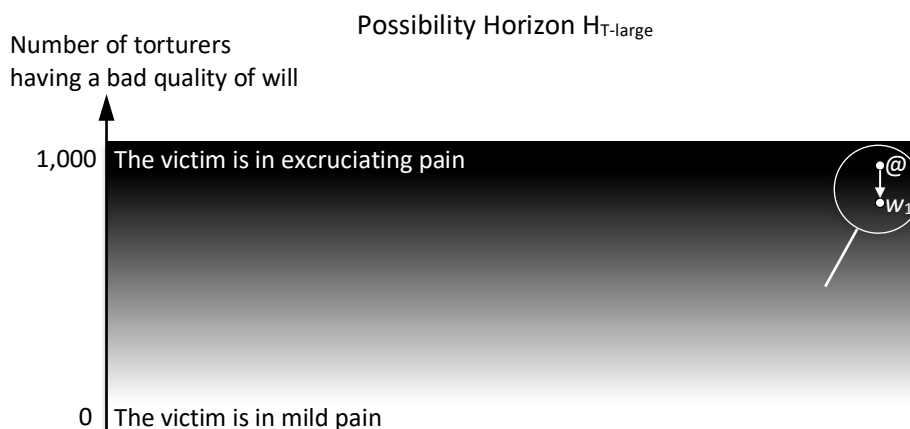
---

<sup>1</sup> In "Making a Vague Difference" (Chapter 8), I used a slightly different version of HARMLESS TORTURERS where the victim is not in pain before any switch is flipped. I used that version since Kagan (2011) uses it in his argument. He argues that there must be a first pain state where the victim is in pain. In Parfit's (1984) original version, there are a thousand torturers, and a thousand victims.

<sup>2</sup> See THE UP-IN-THE-AIR CONDITION on p. 153.

torturer caused the victim's being in such pain. These conclusions seem strange. It seems that what the torturers did had something to do with the victim's being in pain. Therefore, we have reasons to scrutinise the situation more closely. One important way of doing so is to look for more possibilities to include in our possibility horizon.

For another thing, the smaller possibility horizons that the torturers appeal to in their defences do not fit together. If we address one torturer alone, his defence might seem persuasive. However, his defence presumes that we treat it as a live possibility that he had had a different quality of will, while we treat the quality of will of the other torturers as fixed. So, if we simultaneously consider the defences of all torturers, we have to presume both that each torturer could have had a different quality of will, and that he or she could not. This is contradictory. We cannot at the same time consider any given torturer's quality of will as both a background condition and as a candidate cause. Therefore, we should consider a larger possibility horizon, treating each torturer's quality of will as a candidate cause. I could not possibly depict all of these worlds individually, but the illustration below will hopefully do:



Here, @ is the actual world where all the torturers flip their switches, and  $w_1$  is a world (or a range of worlds) where all but one torturer flip their switches.

Given this larger possibility horizon, BLAMEWORTHINESS FOR entails that each torturer is blameworthy for the victim's being in excruciating pain rather than in mild pain. For one thing, focusing on this larger possibility horizon lets us consider the contrast between the victim's being in mild pain and the victim's being in excruciating pain, rather than merely on the difference that is made by flipping a single switch. This larger contrast is crucial for satisfying the first condition of BLAMEWORTHINESS FOR, without relying on controversial assumptions about the

moral relevance of imperceptible differences. Clearly, (i) the victim's being in excruciating pain *just is* worse than the victim's being in mild pain.

Furthermore, (ii) given this larger possibility horizon, each torturer's having a bad quality of will towards the victim rather than the required one is a non-deviant cause of the victim's being in excruciating pain rather than in mild pain. Consider any torturer. (a) This torturer's bad quality of will at  $t$  is process-connected to the victim's being in excruciating pain. His having a poor quality of will belongs to a set of events that is time-sensitively sufficient for the victim's being in excruciating pain. More elaboratively, his poor quality of will belongs to a set of events that guarantees (together with the laws of nature) that the victim will be in excruciating pain at all, that the victim will be in excruciating pain roughly at the time at which he is, and that the victim will be in excruciating pain exactly at the time at which he is. And, if this torturer instead had had the required quality of will, this set would no longer guarantee this. Moreover, the process-connection remains even if we add more and more intermediate times between  $t$  and the time at which the victim is in excruciating pain (such as the time at which this torturer flips his switch).

(b) The victim's being in excruciating pain is less secure in the closest-to-@-at- $t$  world where this particular torturer has the required quality of will. Had this torturer had the required quality of will, one thing less would need to change in order for the victim not to be in excruciating pain. For a similar reason, the victim's being in mild pain is more secure in the closest-to-@-at- $t$  world where this torturer has the required quality of will. All this is true of each torturer.

So, each torturer' having a bad quality of will towards the victim rather than the required one is a cause of the victim's being in excruciating pain rather than in mild pain. Further, there is no reason to believe that this is a case of deviant causation. The outcome is brought about roughly as the torturers planned, and there is no reason to think that causation went astray before they formed their intentions to flip the switches. Therefore, all conditions of BLAMEWORTHINESS FOR are satisfied. This principle entails that each torturer is blameworthy for the victim's being in excruciating pain rather than in mild pain.

You might doubt that each torturer's having a bad quality of will rather than the required one is a cause of the victim's being in excruciating pain rather than mild pain since there is no sharp threshold between being in excruciating pain and being in mild pain. How could a set of events then be minimally sufficient for the victim's being in excruciating pain? And how could we assess the security of the victim's being in excruciating pain rather than in mild pain if these events do not have exact conditions of occurrence?

This objection rests on a mistaken understanding of vagueness. The reason it is vague when the victim starts feeling excruciating pain is not that there is this one event, the victim's feeling excruciating pain, with imprecise conditions of occurrence. Rather, there are many different events, each with precise conditions of

occurrence, but we have not made up our minds about which of these events is the correct referent of “the victim’s being in excruciating pain” (see Fine 1975; Lewis 1986c). According to an epistemic view of vagueness, there is one correct referent, we just do not know which one (see Sorensen 1988; Williamson 1994). According to supervenience, there are several admissible sharpenings of “the victim’s being in excruciating pain”, each referring to a different event with precise conditions of occurrence (see Fine 1975; Keefe 2000). Either way, there is a threshold for when the victim starts being in excruciating pain. This means that flipping one switch belongs to a set of events that is time-sensitively sufficient for bringing about the precise event that did occur. This also means that flipping a switch does make the victim’s being in excruciating pain more secure: there is a precise distance between the actual world and the world where the victim starts being in excruciating pain.<sup>3</sup>

This strategy of individuating events is not open to the expected utility approach (or any approach that take making a difference for whether some outcome occurs to be the only thing that matters for which outcomes you are blameworthy for). Even though there are precise conditions for when an event occurs, this does not necessarily mean that it is morally relevant that one event occurs rather than another. For instance, on the assumption that imperceptible differences in pain are morally irrelevant, it does not matter whether, for instance, 900 or 901 torturers flip their switches in HARMLESS TORTURERS. Even though 900 flipped switches result in another event than 901 flipped switches (some extra current runs through the body of the victim), the difference between these two events is not morally irrelevant. The morally relevant differences emerge when you contrast events that are further away from each other, such as the event when no torturer flips his switch and the event when all do. BLAMEWORTHINESS FOR gives the right verdict in non-threshold cases even if there is no perceptible threshold in harm, and even if imperceptible harms are morally irrelevant since it lets us take such differences into account.

## Climate Change

Does BLAMEWORTHINESS FOR give the intuitively correct verdict about who is blameworthy for climate change and its related harms? I think it does. Climate change has been claimed to be a case of overdetermination (Cripps 2013), a pre-emption case (Lawford-Smith 2016; Eriksson 2019), a collective impact case with a threshold (Kagan 2011) a collective impact case without a threshold (Nefsky 2012; Kingston & Sinnott-Armstrong 2018; Nefsky 2019),<sup>4</sup> and a case where each act does

---

<sup>3</sup> This strategy is explained in further detail in “Reasons for Action” (Chapter 5).

<sup>4</sup> To be more precise, Nefsky does not say that climate change is a non-threshold case. Rather, she says that we cannot exclude that it is.

make a difference for the outcome (Broome 2019). The short answer for why BLAMEWORTHINESS FOR gives the intuitively right verdict about blameworthiness in the case of climate change is that it gives intuitively right verdicts in all these kinds of cases. I made a similar point in relation to REASON.

Moreover, BLAMEWORTHINESS FOR can explain our torn intuitions about whether I am blameworthy for climate change and its related harms if I go joy-guzzling without any regard for the climate. For illustration, say that climate change is a non-threshold case. There is no threshold such that if  $n$  drives with a fossil fuel driven car occur, some climate change related harm will occur, but if  $n + 1$  such drives occur, this harm will occur. If we treat others' quality of will as fixed, we then get a small possibility horizon according to which there is no possibility that my quality of will in relation to climate change makes a morally relevant difference for climate change and its related harms. If we also take this smaller possibility horizon to be the one that is relevant for assessing blameworthiness in this case, we will get the result that I am not blameworthy for climate change and its related harms if I joy-guzzle. It is not up in the air whether any climate-change-related harm will occur. Whether or not I have a substandard quality of will towards the climate does not matter. Sure, it might reveal something reproachable about my character, but it does not make me blameworthy for climate change and its related harms.

However, if we do not treat others' quality of will as fixed, we will get a much larger possibility horizon according to which there is a possibility that severe climate change and its related harms will be severe, and a possibility that these severe harms will be avoided. According to this larger possibility horizon, I am blameworthy for joy-guzzling out of disregard for the climate. (i) I have a climate-change-related reason to refrain from joy-guzzling rather than to joy-guzzle (see also discussion in Chapter 6). That is, my going joy-guzzling is worse than my not going joy-guzzling in virtue of severe future climate-change-related harms being worse than not severe future climate-change-related harms. (ii) there is a time  $t$  such that my having a poor quality of will towards climate change rather than not is a non-deviant cause of my going joy-guzzling rather than not within this larger possibility horizon. If I had had the required quality of will, I had refrained from joy-guzzling,<sup>5</sup> so my having a poor quality of will towards the climate is a cause of my going joy-guzzling rather than not. Further, we have no reasons to think that my disregard for the climate *deviantly* caused me to joy-guzzle. So, BLAMEWORTHY FOR entails that I am blameworthy for joy-guzzling rather than not given this larger possibility horizon.

As I argued in Chapter 6, this larger possibility horizon is the relevant one. The smaller possibility horizon leaves important issues unexplained and requires us to

---

<sup>5</sup> This partly follows from the WHETHER-WHETHER INFERENCE (p. 138), and partly from the fact that my having a poor quality of will towards the climate is process-connected to the outcome.

make contradictory assumptions about whether there was a relevant possibility that others could have cared as required about the climate.

As a final point, *BLAMEWORTHINESS FOR* does not only entail that I am blameworthy for joy-guzzling, but also that I am blameworthy for climate change being severe rather than not. That climate change is severe just is worse than its not being severe, and if my having a substandard quality of will rather than not is a non-deviant cause of climate change being severe rather than not (which it is if I for instance go joy-guzzling once), I am blameworthy for climate change being severe rather than not, or at least that I will be when these harms occur.

You might think that this seems excessive. Why would I be blameworthy for the entirety of future climate change and its related harms just because I uncaringly go joy-guzzling once? If you think so, I want to remind you that *BLAMEWORTHINESS FOR* does not say anything about degree of blameworthiness. It is an open question whether others are warranted in reacting very negatively to you in virtue of being blameworthy for climate change, or just mildly so. They might for instance be warranted in being outraged for what you have done, or just mildly annoyed about what you have done. Plausibly, your degree of blameworthiness for some climate change related harm depends on how much you have contributed to this harm. If your contribution is small, you are less blameworthy than if your contribution is large. That is, if your having a poor quality of will rather than the required one only slightly increased the security of the bad outcome, you are less blameworthy than if your having a poor quality of will rather than the required one hugely increased the security of this outcome. However, to give a more accurate account of degree of blameworthiness has to be the topic for another day.

## Penned-In Sharks

Fischer and Ravizza's (1998) influential account of moral responsibility produces intuitively correct verdicts about who is morally responsible for what in a wide range of cases. Still, as they to some extent concede, it also sometimes gives counterintuitive verdicts. In this section, I will first present Fischer and Ravizza's account and explain why this is the case. I will then proceed to showing that *BLAMEWORTHINESS FOR* gives the desired verdicts in the cases where Fischer and Ravizza's account fails to do so.

Fischer and Ravizza (1998) suggest that one is morally responsible for X, where X might be an action, omission or outcome, only if one had guidance control over X.<sup>6</sup> One has guidance control over an action or omission if and only if this action or

---

<sup>6</sup> Besides the guidance control condition, they say that there also is an epistemic condition for being morally responsible.

omission “issues from one’s own, moderately reasons-responsive mechanism” (133). A mechanism consists of “the process that leads to the relevant upshot” (38). A characteristic process leading to action is when someone first deliberates about what to do, and then decides upon a course of action. Fischer and Ravizza do not present any exact principle for how to individuate different mechanisms. Instead, they “rely on the fact that people have intuitions about fairly clear cases of ‘same kind of mechanism’ and ‘different kind of mechanism.’” (40). For instance, they rely on the intuitive judgement that normal deliberation is a different kind of process – a different mechanism – than deliberation induced by hypnosis, irresistible urges, significant electronic manipulation of the brain, drugs, or something of the kind.

Making a mechanism one’s own is in large part a matter of accepting that one is properly held responsible for actions and omissions resulting from processes of that type. One might for example accept that one is responsible for actions and omissions that issues from one’s normal reasoning, but not that one is morally responsible for actions and omissions that issues from irresistible urges.

The required reasons-responsiveness of the mechanism is, in turn, a matter of whether the actual process – the mechanism – leading to the action or omission is suitably *receptive* and *reactive* to reasons for acting otherwise. The mechanism is suitably receptive if there an understandable pattern of scenarios – actual and hypothetical – where the agent recognises that there are sufficient reasons to act otherwise; and it is suitably reactive if the agent also acts otherwise in at least one such scenario on the basis of the relevant recognition. To evaluate whether the mechanism is suitably receptive and reactive to reasons to act otherwise, Fischer and Ravizza tell us to hold fixed the operation of the actual mechanism. Thus, in cases where the mechanism actually issuing in action has not been subjected to hypnosis, brainwashing, coercion or any other form of manipulation, we should not evaluate whether the mechanism is receptive or reactive to reasons by considering scenarios where the mechanism has been subjected to such influences.

Fischer and Ravizza show that their account yields correct verdicts in Frankfurt-style cases such as the following (here presented in shortened form):

ASSASSIN: Sam confides in Jack that he wants to kill the mayor. Sam is disturbed about the mayor’s liberal policies. Whereas Sam’s reasons for killing the mayor are bad, they are *his* reasons: he has not been hypnotised, brainwashed, coerced, and so forth. Jack is pleased with Sam’s plan, but to make sure that Sam will go through with it, he secretly installs a device in Sam’s brain, which allows him to monitor Sam’s brain activity, and to intervene in it, if he desires. If Sam would show any sign of wavering, Jack will intervene and make sure that Sam carries out his original plan. As things turn out, Sam does not show any sign of wavering, and shoots the mayor as a result of his original deliberations. Jack does not intervene.

(Fischer & Ravizza 1998: 29)

Fischer and Ravizza claim that Sam is blameworthy for shooting the mayor. I agree. Further, as they point out, Sam is blameworthy for shooting the mayor even though he could not have acted otherwise. Jack – the counterfactual intervener – ensures that Sam will shoot the mayor. So, ASSASSIN is a counterexample to the idea that you can be blameworthy for what you do only if you could have acted otherwise. That is, ASSASSIN is a counterexample to the principle of alternate possibilities (PAP).<sup>7</sup> (This is what makes ASSASSIN a Frankfurt-style case.)<sup>8</sup>

Even though Sam could not have acted otherwise, he still has guidance control over his actions. The process leading to his shooting the mayor is suitably reasons-responsive. To decide this, we have to hold fixed the actually operating mechanism issuing in action. This means that we have to hold fixed the non-intervention of Jack: In scenarios where Jack intervenes, the mechanism that issues in Sam's shooting the mayor is relevantly different from the one issuing in Sam's action in the actual world. If we hold the actually operating mechanism fixed, we find that there is an understandable pattern of alternative scenarios where Sam recognises that there is sufficient reason to act otherwise. The mayor might for instance change her policies before Sam carries out his plan; or she might not have had liberal policies to begin with. In such scenarios, we can readily assume, Sam would have recognised that he has sufficient reasons not to shoot the mayor. Therefore, the actually operating mechanism is suitably receptive.

The actually operating mechanism is also suitably reactive. We might assume that Sam does not end up shooting the mayor in some of the scenarios where he recognises that he has sufficient reasons not to do so. Nothing in the description of the case indicates otherwise. For example, in some scenarios where the mayor changes her policies and Sam as a result recognises that he has sufficient reasons to abandon his plans, he also abandons his plans.

Finally, the mechanism actually issuing in Sam's killing the mayor is relevantly his. He was not hypnotised, brainwashed, or anything like that, but shoots the mayor as a result of his own original deliberations. Even though the example does not say, we might assume that Sam accepts that he is properly held responsible for actions and omissions resulting from processes of that type.<sup>9</sup> We can thus conclude that Sam's shooting the mayor issues from his own, moderately reasons-responsive mechanism.

To have guidance control over an *outcome* is a more complex affair. One has guidance control over an outcome O if and only if (I) one has guidance control over

---

<sup>7</sup> This principle was introduced on p. 236.

<sup>8</sup> Note that ASSASSIN is not the same case as ASSASSINS, introduced on p. 24.

<sup>9</sup> For comparison, if Jack had intervened, the process leading to Sam's shooting the mayor would not have been Sam's. Likely, Sam would not accept that he is properly held responsible for decisions and actions that others have induced in him via an implanted device.



the action or omission that brings about O, and (II) the outcome O is suitably sensitive to one's failure to act otherwise.<sup>10</sup> I have already described how you decide whether one has guidance control over an action or omission. The procedure for deciding whether O is suitably sensitive to one's failure to act otherwise reminds of the procedure for deciding whether a mechanism is suitably receptive and reactive. To decide whether O is suitably sensitive to one's failure to act otherwise, Fischer and Ravizza tell us to hold fixed all triggering events that occurs after one's failure to act otherwise but before the outcome occurs, and then consider whether – given this restriction – the outcome still would have occurred if one had acted otherwise.<sup>11</sup>

A triggering event is “an event which is such that, if it were to occur, it would *initiate* a causal sequence leading to [the relevant outcome]” (110-11). They do not suggest any exact way of deciding what it is for an event to initiate a causal sequence, but rely on the fact that we have fairly clear intuitions about what it is for an event to initiate such a sequence. For instance, if John were to intervene in ASSASSIN, making Sam arrive at the decision to shoot the mayor and also follow through on this decision, John would initiate a causal sequence leading to the mayor's being shot. BACKUP BILLY provides another example. In this case, Suzy throws a rock at a window, breaking it, while Billy is lurking in the background, ready to throw his rock in case Suzy would not have thrown her.<sup>12</sup> Fischer and Ravizza do not discuss this case, but it seems clear that Billy's throwing the rock is a triggering event. If this event were to occur, it would initiate a causal sequence leading to the window breaking.

Fischer and Ravizza propose several cases to illustrate the idea that one has guidance control over an outcome if and only if one satisfies (I) and (II), including the following early pre-emption case (which I have abridged):

MISSILE 2: An evil woman, Elizabeth, has obtained a missile and missile launcher, and she has decided (for her own rather perverse reasons) to launch the missile toward Washington, D.C. Suppose that Elizabeth's situation is like that of Sam; she has not

---

<sup>10</sup> I have here simplified Fischer and Ravizza's account somewhat. They say that one has guidance control over an outcome O if and only if (I) one has guidance control over *one's bodily movements that constitutes the action or omission* that brings about O, and (II) the outcome O must in turn be suitably sensitive to one's *failure to move one's body* in a certain alternative way. For our purposes here, the simplification I make in the main text is of no importance.

<sup>11</sup> Again, I have simplified Fischer and Ravizza's (1998) account. They say that an outcome is suitably sensitive if and only if (i) it is the case that if the agent had moved his body in some alternative way *B\** at the relevant time *T*, (ii) all other triggering events (apart from *B\**) which do *not actually* occur between *T* and *T+i* were *not* to occur, and (iii) a *P*-type process were to occur, then *O* would not occur. (see Fischer & Ravizza 1998: 112). Again, this simplification is of no importance for our purposes. The arguments I make do not depend on questions about how do individuate different causal processes.

<sup>12</sup> This case was introduced in Chapter 11, on p. 222.

been manipulated, brainwashed, and so forth. Further, there is another woman, Carla, who would launch the missile if Elizabeth were to refrain. As things turn out, Elizabeth (and not Carla) launches the missile, and Washington D.C. is bombed.

(Fischer & Ravizza 1998: 93-94)

Intuitively, Elizabeth is morally responsible for bombing of Washington D.C. This is also what Fischer and Ravizza's account entails. (I) Elizabeth has guidance control over her actions when launching the missile. Her launching the missile issues from her own, moderately reasons-responsive mechanism.<sup>13</sup> Further, (II) the bombing of Washington D.C. is suitably sensitive to her failure to refrain from launching the missile. To see this, we have to hold fixed the non-occurrence of all triggering events. Carla's launching the missile is such an event. If this event were to occur, it would initiate a causal sequence that would result in the bombing of Washington D.C. Given that we hold fixed the non-intervention of Carla, we find that the bombing of Washington had not occurred if Elizabeth had refrained from launching the missile. So, both (I) and (II) are satisfied.

We could replicate this line of reasoning, showing what Fischer and Ravizza's account not only gives the intuitively correct verdict in Frankfurt-style cases, but also in early pre-emption cases.

There are however cases where their account produces counterintuitive verdicts. Fischer and Ravizza (1998) considers one such case, originally proposed to them by David Kaplan:

[PENNED-IN SHARKS:] John is walking along a beach, and he sees a child struggling in the water. John believes that he could save the child with very little effort, but he is disinclined to expend any energy to help anyone else. [... Unbeknownst to John], a bad man wants to make sure that the child (struggling in the water) is not saved. He has penned in a number of hungry sharks, which he will release if and only if John were to jump into the water. As it happens, John does not jump into the water [...], and thus the bad man keeps the sharks in their pen; but had John jumped in, the bad man would have released the sharks, and they would have eaten John.

(Fischer & Ravizza 1998: 125; 138)

Intuitively, John is not morally responsible for the death of the child. Because of the presence of the penned-in sharks, there is no possibility that he will save the child even if he tries. However, Fischer and Ravizza's account indicates that John is morally responsible for the death of the child. It implies that the death of the child

---

<sup>13</sup> I have omitted to show that Fischer and Ravizza's account entails that Elisabeth has guidance control over her actions when launching the missile. It does, but I will leave it up to you to check this.

is suitably sensitive to John's failure to try to save the child. If we hold fixed the fact that the bad man does not release the sharks (a triggering event), we find that the death of the child counterfactually depends on John's actions. The child would not have drowned if John had tried to save the child. Therefore, (II) is satisfied. Moreover, (I) is also satisfied. John's continued stroll by the beach issues from his own, moderately reasons-responsible mechanism.<sup>14</sup>

Fischer and Ravizza (1998) admits that their account indicates that John is morally responsible for the death of the child, but argue that they are justified in thinking that their account gives the right verdict in this case. They say that cases like PENNED-IN SHARKS are "puzzling and difficult" (138), and argue that since their account gives correct verdicts in other similar cases (like ASSASSIN and MISSILE 2) where our intuitions are clearer, we have strong reasons to believe that their account gives the right verdict also in PENNED-IN SHARKS, where our intuitions are less clear. As they say, "if this way of treating such cases is indeed correct, then we submit that our treatment of 'Pinned-In Sharks' is *also* correct" (139).

Several writers have been dissatisfied with this response. Carl Ginet (2006) says that he does not agree with Fischer and Ravizza's intuition that John is morally responsible for the death of the child, and Randolph Clarke (2011) writes that "Fischer and Ravizza are surely wrong about PENNED-IN SHARKS" (609).<sup>15</sup>

In a response to Ginet, Fischer (2006) points out that he and Ravizza did not claim that the intuitive verdict about PENNED-IN SHARKS is that John is morally responsible for the death of the child, but rather that whether John is morally responsible for the death of the child is a difficult matter, and given that their account gives the intuitive verdict in other similar cases, they have reason to believe that their account gives the right verdict also in this case where our intuitions are less clear.

I agree with Fischer (2006) that we might have reasons to accept the verdicts our theories yield even if these verdicts at first sight seem odd. We might have such reasons if our theories give the right verdicts in a range of similar cases, and if there are no other theories that can accommodate all our intuitions. That said, I agree with

---

<sup>14</sup> There is an understandable pattern of alternative scenarios where John recognises that there are sufficient reasons to save the child. There might for instance be a large revenue for anyone who saves a drowning child at the beach. In such a case, John might think he has sufficient reasons to jump in the water and try to save the child. So, the actually operating mechanism issuing in John's continued stroll is suitably receptive to reasons. Further, in some such scenarios, John might jump into the water and try to save the child. Hence, the actually operating mechanism is also suitably reactive. Finally, the process leading John to continue his walk along the beach in the actual scenario is relevantly his. He is for instance not hypnotised, brainwashed or otherwise manipulated into continuing his stroll, so there is no reason to think that he does not accept being held morally responsible for the type of mechanism actually issuing in action.

<sup>15</sup> Clarke suggests several other counterexamples similar to PENNED-IN SHARKS, effectively showing that PENNED-IN SHARKS is not just one isolated case which we can ignore.

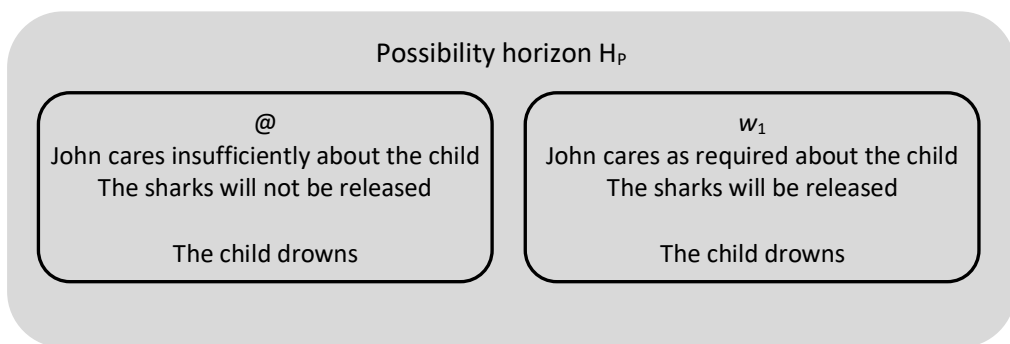
Ginet and Clarke that our intuitions are clear in PENNED-IN SHARKS. Surely, John is not morally responsible for the death of the child. As far as I can see, PENNED-IN SHARKS is puzzling and difficult only to the extent that you believe in a theory that implies that John *is* morally responsible for the child's death. That Fischer and Ravizza's account implies that John is morally responsible for the death of the child is a problem for their account. If we could find an account that, like theirs, give the right verdicts in Frankfurt-style cases like ASSASSIN and in early pre-emption cases like MISSILE 2, and that also gives the right verdict in PENNED-IN SHARKS, we should prefer that theory.

What kind of theory are we looking for? Fischer and Ravizza's strategy gives the wrong verdict about PENNED-IN SHARKS since they insist that we hold the non-occurrence of all triggering events fixed when evaluating whether the child's death was suitably sensitive to John's omission. When we hold fixed the fact that the sharks were not released (as in the actual world), we overlook the relevant possibility that the sharks will be released in the counterfactual world where John jumps into the water. Still, if we give up the idea that we should hold fixed the non-occurrence of all triggering events, their account no longer yields the right verdicts in early pre-emption cases. In MISSILE 2, for example, if we do not hold fixed the non-occurrence of Carla's intervention, we find that the bombing of Washington D.C. was not suitably sensitive to Elisabeth's failure to refrain from launching the missile: the bombing would have occurred even if Elisabeth had refrained from launching it. By extension, we would also find that Elisabeth lacked guidance control over the bombings, and so that she is not morally responsible for this outcome. This would surely be the wrong verdict. So, we need an account that let us take into account all relevant possibilities in cases like PENNED-IN SHARKS, but that is sensitive enough to yield the verdict that the bombing of Washington D.C. occurred because Elisabeth launched the missile.

BLAMEWORTHINESS FOR is such an account. We have already seen that it can deliver the right verdict in early pre-emption cases (in Chapter 11). In brief, it entails that Elisabeth is blameworthy for the bombing of Washington D.C. because (i) it is bad that Washington D.C. is bombed, and (ii) Elisabeth's bad quality of will towards the bombing is a non-deviant cause of it. CAUSATION helps us see that condition (ii) is satisfied. (a) Elisabeth's bad quality of will is process-connected to the bombing. It belongs to a set of events that guarantees, given the laws of nature, that the bombing will occur, and this set does no longer guarantee, given the laws of nature, that the bombing will occur if we remove Elisabeth's bad quality of will from this set. And we get no reason to re-evaluate this verdict if we consider more and more temporally fragile versions of bombing, or if we consider more and more intermediate times between the relevant time at which Elisabeth had a bad quality of will towards the bombing and the bombing. Moreover, (b) her having a bad quality of will made this outcome more secure. In the actual world where she launches the missile, two things would need to change in order for Washington D.C.

not to be bombed: Elisabeth’s quality of will and Carla’s quality of will. Whereas in the closest relevant possible world where Elisabeth has the required quality of will, only one thing needs to change in order for Washington D.C. not to be bombed: Carla’s quality of will. Further, since we have no reasons to believe that MISSILE 2 is a case of deviant causation, we can conclude that BLAMEWORTHINESS FOR entails that Elisabeth is blameworthy for the bombing of Washington D.C.

BLAMEWORTHINESS FOR also delivers the right verdict about PENNED-IN SHARKS. To see this properly, we first have to settle the relevant possibility horizon. Normally, we should treat it as an open possibility that each agent who has a substandard quality of will towards the child’s survival could have had the minimally required quality of will. This is what RELEVANT POSSIBILITIES FOR BLAME tells us. So, normally, we should treat it as an open possibility that John had cared as required about the child, and that the bad man had done the same. However, in PENNED-IN SHARKS, there is no possibility that the bad man had refrained from releasing the sharks in case John jumps into the water. As the case is presented, he “will release [the sharks] if and only if John were to jump into the water” (138). This strongly suggests that there is no possibility that the bad man had cared as required about the child. So, we end up with only two relevant possibilities: either John cares insufficiently about the child’s survival, with the result that he does not jump into the water, the sharks are not released, and the child drowns; or, he cares as required about the child’s survival, with the result that he gets into the water, the sharks are released, John gets eaten alive, and the child drowns.



As you see, the child drowns in all relevant possible worlds. Whether John cares as required or not about the child’s survival makes no difference for whether the child drowns. It makes a difference only to how this outcome will come about. In this respect, John’s quality of will is a mere switch, and PENNED-IN SHARKS is a switching case. In such cases, we usually deem that the potential cause (John’s quality of will in this case) upon closer reflection is not a cause. This is also what CAUSATION entails: John’s having a poor quality of will towards the child’s survival

rather than the required one is not a cause of the child's drowning rather than surviving. Even though John's poor quality of will is process-connected to the child's drowning, the child's drowning is not more secure in the closest-to-@-at-*t* world where John has a poor quality of will towards the child (which is @) than it is in the closest-to-@-at-*t* world where John has the required quality of will ( $w_1$ ). That is, condition (b) of CAUSATION is not satisfied.<sup>16</sup>

Since John's having a poor quality of will rather than the required one is not a cause of the child's drowning rather than surviving, BLAMEWORTHINESS FOR entails that John is not blameworthy for the child's drowning rather than surviving. Condition (ii) of this principle is not satisfied.

With this, I submit that BLAMEWORTHINESS FOR gives a better explanation for our intuitions about blameworthiness than Fischer and Ravizza's account does. It gives the right verdict in Frankfurt-style cases like ASSASSIN,<sup>17</sup> in early pre-emption cases like MISSILE 2, and on top of that, it gives the right verdict in PENNED-IN SHARKS.

One last thing. It seems to me that Fischer and Ravizza's strategy will produce counterintuitive verdicts in late pre-emption cases. As they define triggering events, they are events that – were they to occur – they would initiate a causal sequence leading to the relevant outcome. In a case like BOTTLE SHATTERING (p. 81), where both Suzy and Billy throw their rocks but where Suzy's rock hits the bottle first, it turns out that Billy's throwing his rock is not a triggering event. It is not an event that, were it to occur, it would initiate a causal sequence leading to the bottle shattering. It did occur, and it did not initiate such a sequence. So, Fischer and Ravizza give us no reason to think that we should hold fixed the non-occurrence of Billy's throwing his rock when evaluating whether the bottle shattering was suitably sensitive to Suzy's throwing her rock. As a result, it turns out that the bottle shattering was not suitably sensitive to Suzy's throwing her rock. The bottle shattering would occur whether Suzy throws her rock or not. So, on their account, it turns out that Suzy does not have guidance control over the bottle shattering, and that she therefore cannot be morally responsible for it. This verdict seems wrong. If my line of reasoning here is correct, late pre-emption cases pose an additional problem for Fischer and Ravizza's account. BLAMEWORTHINESS FOR, however, gives intuitively correct verdicts in late pre-emption cases.

---

<sup>16</sup> For more on how BLAMEWORTHINESS FOR applies to switching cases, see Chapter 11.

<sup>17</sup> I have not showed that BLAMEWORTHINESS FOR gives the right verdict about ASSASSIN here. Instead, I refer back to the discussion about BACKUP NEUROSCIENTIST BILLY in Chapter 11, an example that is structurally the same as ASSASSIN.

## A Potential Counterexample

The idea that a cause increases the security of its effect is in some ways similar to the idea that a cause raises the probability of its effect.<sup>18</sup> Because of this similarity, counterexamples to the idea that causes always raises the probability of its effect are also potential counterexamples to the idea that causes always increase the security of its effect. Consider for instance the following example:

UNSKILLED SUZY: Suzy is walking down the street. When she reaches the big house on the corner, she stops and considers. She has an intense dislike for the elderly couple who live in the house, and she has just had an idea: she is going to upset them by breaking their window on the first floor. She carefully selects a rock and hurls it towards the window. She is terrible at throwing, and just barely manages to hit her target. A small gust of wind would have made her miss it. She feels a jolt of satisfaction when she hears the sound of breaking glass. Unbeknownst to Suzy, Billy is lurking in the background. On seeing that Suzy throws her rock, Billy is satisfied and walks away. However, if Suzy had not thrown her rock, Billy, who is an excellent thrower, would have thrown a rock himself a moment later, breaking the window.<sup>19</sup>

You might think that this case shows that an event might be a cause even if it makes the outcome less secure, and so that this case is a counterexample to CAUSATION and BLAMEWORTHINESS FOR. You might argue that, in this case, Suzy's throw clearly causes the window to break. Still, Suzy's throwing her rock makes the window breaking less secure. She is terrible at throwing – had there been a small gust of wind, she would have missed – and had she not thrown her rock, the excellent thrower Billy would have thrown his rock instead, and he would have hit the window. So, this seems to be a case where an event causes another without making it more secure. Should we therefore give up the requirement that a cause must make an outcome more secure? Should we let go of CAUSATION's condition (b)?

We should not. For one thing, process-connections are not sufficient for causation. (And the same goes for any account of causation that builds on minimal sufficiency, such as NESS.) We added the requirement that a cause must increase the security of its outcome for a reason. Without this requirement, we get wrong verdicts about

---

<sup>18</sup> Lewis (1986a), Menzies (1989) and Fenton-Glynn (2017) propose accounts of this kind..

<sup>19</sup> Menzies (1989), Schaffer (2001) and Glynn (2011) discuss similar examples in relation to theories of probabilistic causation. The standard counterexample to theories of probabilistic causation goes something like this: Billy throws the rock, raising the probability of the window breaking, but Suzy's rock hits the window first. Here, it seems that Suzy's throw (not Billy's) caused the window to break. However, Billy's throw raised the probability of the window breaking, and so simple probabilistic accounts of causation mistakenly entail that also his throw is a cause of the window breaking. These accounts could avoid giving this verdict if they also required that a cause must be process-connected to its effect. Events that raise the probability of an outcome without causing it are commonly called fizzlers. For discussion, see also Touborg (2018: 239-43).

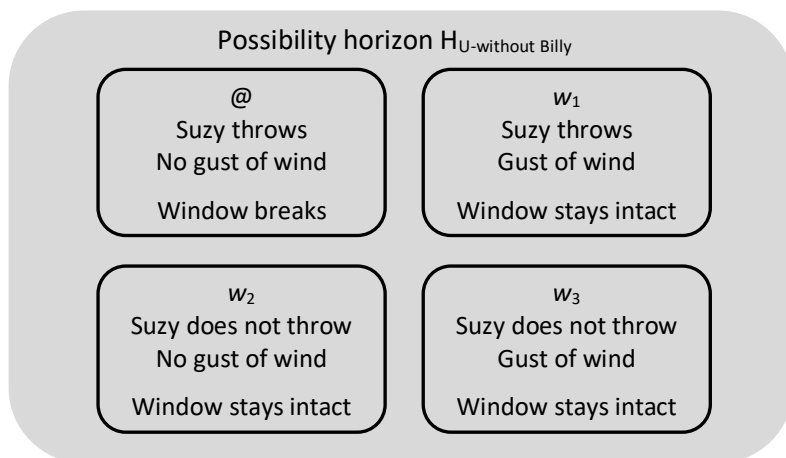
causation and blameworthiness in switching cases (like TROLLEY TROUBLE) and in cases of causal transitivity failure (like CAR KEYS). We also cannot accommodate contrastive causal claims, and we lose our way to distinguish causes and background conditions.<sup>20</sup> So, if we let go of condition (b), we are in a lot of trouble.

Moreover, it is far from sure that UNSKILLED SUZY really is a counterexample to CAUSATION and BLAMEWORTHINESS FOR. You could view the situation in UNSKILLED SUSY in at least three different ways, each with some intuitive appeal, and none of them is a counterexample to CAUSATION and BLAMEWORTHINESS FOR. I will here go through these ways of understanding the situation.

### Disregarding Billy

First, it might seem legitimate to disregard Billy’s readiness to intervene because, as things turned out, he did not do anything. He just stands there in the background, observing. (To disregard Billy’s readiness to intervene concur with Fischer and Ravizza’s (1998) advice to hold fixed the actual behaviour of counterfactual interveners and other triggering events.) Still, if we disregard the fact that Billy is ready to intervene, it seems that Suzy’s throw caused the window breaking. If we disregard Billy, the window breaking counterfactually depends on Suzy’s throw. Had Suzy not thrown her rock, the window breaking would not have occurred. Moreover, if we remove Billy from the picture, it seems that Suzy is blameworthy for breaking the window. Just out of spite, she broke the elderly couple’s window.

Interestingly, this is also what CAUSATION and BLAMEWORTHINESS FOR entails. Consider first CAUSATION. If we remove Billy from the picture, there are four relevant possibilities at the time when Suzy throws her rock, as follows:



<sup>20</sup> See e.g. discussion in Chapter 11, p. 220ff, and in Chapter 12, p. 254ff.



This is a standard case where two events are independently necessary and jointly sufficient for the outcome. The window breaking counterfactually depends on Suzy's throwing the rock, and so CAUSATION entails that Suzy's throwing the rock is a cause of the window breaking.<sup>21</sup>

Using a slightly different possibility horizon focusing on Suzy's quality of will instead of whether she throws the rock, we can also show that BLAMEWORTHINESS FOR entails that Suzy is blameworthy for the window's breaking rather than staying intact, given that we still disregard the fact that Billy is ready to intervene.

Still, this way of understanding the situation in UNSKILLED SUSY does not seem to fully capture what is going on. That Billy is ready to intervene if Suzy does not throw her rock is vital to the story. In addition, RELEVANT POSSIBILITIES FOR BLAME suggests that we should not neglect Billy. It suggests that we should include all agents involved in the situation. So, it seems that  $H_{U\text{-without Billy}}$  is not the relevant possibility horizon in this case. This brings us to the second way in which we could understand UNSKILLED SUZY.

### **The Window Breaking Occurred in Spite of Suzy's Throw**

Second, we might think that the window broke *in spite of* what Suzy did. Suzy is bad at throwing, and had she not thrown her rock, the chances that the window would break would have been even higher. Billy, who is an excellent thrower, would have thrown his rock, and the window breaking would have been even more secure if he did. We might think that if Suzy had known about Billy's presence, skills and intent, she would probably not have thrown her rock, risking to miss the window. Since she wants the elderly couple to get upset, she would probably have waited for Billy to expertly throw his rock instead. If we think of the case like this, it does not seem that Suzy's throw caused the window to break, nor that Suzy is blameworthy for breaking the window. She might be blameworthy for having the intention to upset the elderly couple, and maybe in virtue of being a bad person, but not for the window breaking. The window broke despite what Suzy did.

These are also the verdicts CAUSATION and BLAMEWORTHINESS FOR give. If we do not disregard Billy, there are four things that vary across the relevant possible worlds: Suzy might or might not throw her stone, there might or might not be a gust of wind, Billy might or might not be ready to intervene, and he might or might not succeed if he intervenes. This gives us a possibility horizon containing sixteen possible worlds.

---

<sup>21</sup> Remember THE WHETHER-WHETHER INFERENCE, p. 138. Also, Suzy's throwing the rock is process-connected to the outcome.

Possibility horizon  $H_{U\text{-large}}$

<p>@</p> <p>Suzy throws No gust of wind Billy would intervene Billy would succeed</p> <p>Window breaks</p>	<p><math>w_1</math></p> <p>Suzy throws A gust of wind Billy would intervene Billy would succeed</p> <p>Window stays intact</p>	<p><math>w_2</math></p> <p>Suzy throws No gust of wind Billy would intervene Billy would fail</p> <p>Window breaks</p>	<p><math>w_3</math></p> <p>Suzy throws A gust of wind Billy would intervene Billy would fail</p> <p>Window stays intact</p>
<p><math>w_4</math></p> <p>Suzy throws No gust of wind Billy would not intervene Billy would succeed</p> <p>Window breaks</p>	<p><math>w_5</math></p> <p>Suzy throws A gust of wind Billy would not intervene Billy would succeed</p> <p>Window stays intact</p>	<p><math>w_6</math></p> <p>Suzy throws No gust of wind Billy would not intervene Billy would fail</p> <p>Window breaks</p>	<p><math>w_7</math></p> <p>Suzy throws A gust of wind Billy would not intervene Billy would fail</p> <p>Window stays intact</p>
<p><math>w_8</math></p> <p>Suzy does not throw No gust of wind Billy will intervene Billy will succeed</p> <p>Window breaks</p>	<p><math>w_9</math></p> <p>Suzy does not throw A gust of wind Billy will intervene Billy will succeed</p> <p>Window breaks</p>	<p><math>w_{10}</math></p> <p>Suzy does not throw No gust of wind Billy will intervene Billy will fail</p> <p>Window stays intact</p>	<p><math>w_{11}</math></p> <p>Suzy does not throw A gust of wind Billy will intervene Billy will fail</p> <p>Window stays intact</p>
<p><math>w_{12}</math></p> <p>Suzy does not throw No gust of wind Billy will not intervene Billy would succeed</p> <p>Window stays intact</p>	<p><math>w_{13}</math></p> <p>Suzy does not throw A gust of wind Billy will not intervene Billy would succeed</p> <p>Window stays intact</p>	<p><math>w_{14}</math></p> <p>Suzy does not throw No gust of wind Billy will not intervene Billy would fail</p> <p>Window stays intact</p>	<p><math>w_{15}</math></p> <p>Suzy does not throw A gust of wind Billy will not intervene Billy would fail</p> <p>Window stays intact</p>

Given this possibility horizon, CAUSATION entails that Suzy's throwing the rock rather than not is not a cause of the window's breaking rather than staying intact. Her throw does not make the window breaking more secure (nor the window's staying intact less secure), so condition (b) is not satisfied. To see this, we must compare the security of the window breaking in the closest-to-@-at- $t$  world where Suzy throws the rock (which is @) to the security of the window breaking in the closest-to-@-at- $t$  world where she does not (which is  $w_8$ ). The security of the window breaking in @ is determined by the distance between @ and the closest-to-@-at- $t$  world where the window does not break, which is  $w_1$  where a gust of wind makes Suzy's rock miss the window. This distance is relatively short. There could easily have been a gust of wind that threw Suzy's rock off its course, and Suzy is determined to throw her rock. The security of the window breaking in  $w_8$ , in turn, is

determined by the distance between  $w_8$  and the closest-to- $w_8$ -at- $t$  world(s) where the window does not break. This is either  $w_{10}$  where Billy intervenes but fails to break the window, or  $w_{12}$  where Billy will not intervene, depending on how likely he is to succeed to break the window if he tries, and on how determined he is to intervene. Both these distances are relatively long. Billy is excellent at throwing rocks, and so he is not likely to fail breaking the window if he tries, and he is quite determined (we might imagine) to throw his rock in case Suzy does not throw her. We then see that the window breaking is more secure in @ than it is in  $w_8$ . In @, Suzy might easily have missed her throw because of a gust of wind with the result that the window had stayed intact, but in  $w_8$ , Billy is not likely to fail breaking the window, neither by failing to hit the window nor by abstaining from throwing. So, according to the current possibility horizon, Suzy does not cause the breaking of the window. Instead, the window breaking occurs *in spite of* Suzy's throw.

Using a slightly different possibility horizon, focusing on Suzy's and Billy's quality of wills at a time slightly before  $t$  instead of their doings at  $t$ , we can show that BLAMEWORTHINESS FOR entails that Suzy is not blameworthy for breaking the window, nor for upsetting the elderly couple. She might be blameworthy for having the intention of upsetting the old couple (since this might increase the security of their becoming upset, considering that she is likely to find other some way to upset them if the plan to break the window fails), and she might be blameworthy in virtue of being a bad person, but since her poor quality of will is neither a cause of the window breaking, nor of the elderly couple's becoming upset, she is not blameworthy for these outcomes.<sup>22</sup>

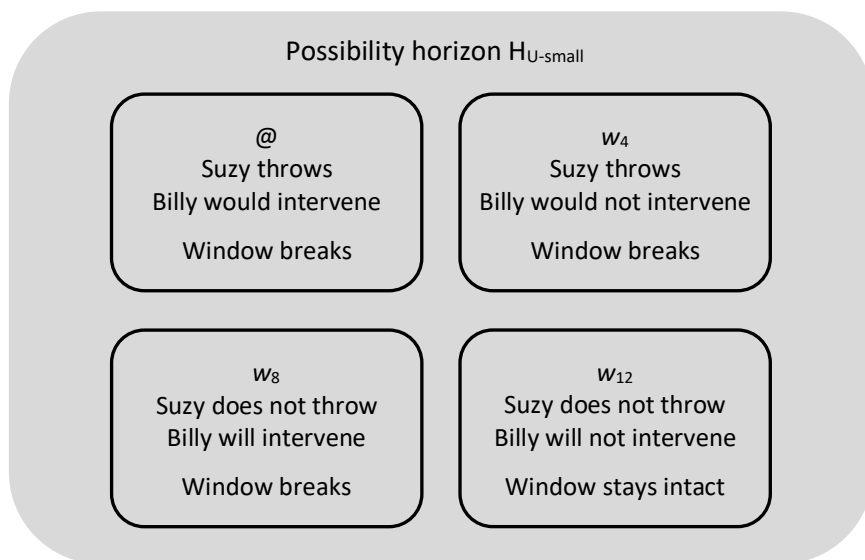
## Suzy Caused the Window Breaking

To say that the window breaking occurred in spite of what Suzy did seems like a plausible story of what is going on in UNSKILLED SUZY. However, we might imagine that the elderly couple would not agree to this story. They might argue that while it seems that the window broke in spite of what Suzy did (since she is such a bad thrower, and since the skilled thrower Billy would have thrown his rock if Suzy had not thrown her), her throw still clearly caused the window breaking, and this is true even if we do not disregard the fact that Billy was lurking in the background. They might argue that, as things actually turned out, there was no gust of wind that made Suzy miss her target, and given that there was no gust of wind or anything else preventing Suzy from hitting the window, she caused the window to break, and is blameworthy for doing so.

---

<sup>22</sup> Remember the distinction between *blameworthiness* and *blameworthiness for* introduced in Chapter 10, and the discussion about being blameworthy for one's intentions towards the end of Chapter 12.

There is something to the elderly couple’s story. They seem to have pinpointed a third way of understanding the situation. Interestingly, CAUSATION and BLAMEWORTHINESS FOR can explain also the elderly couple’s way of understanding the situation. When the elderly couple argues that Suzy caused the window breaking “*given that there was no gust of wind...* “, they are effectively urging us to delimit our possibility horizon, and hold fixed the fact that there was no gust of wind. If we agree to do this, we get a much smaller possibility horizon not treating it as an open possibility that there could be a gust of wind. For the sake of simplicity and symmetry, the following possibility horizon also holds fixed the fact that Billy will succeed in breaking the window if he intervenes. (We would get the same results about whether Suzy’s throw caused the outcome if we treat it as an open possibility that Billy could have missed if he tried breaking the window).



Given this smaller possibility horizon, we find that Suzy’s throw *did* cause the window to break rather than to stay intact. The case has turned into a standard early pre-emption case. First, Suzy’s throw is process-connected to the window breaking. It belongs to a set of events that guarantees, given the laws of nature, that the window breaking will occur, and this set does no longer guarantee, given the laws of nature, that the window breaking will occur if we remove Suzy’s throw from this set. And this picture remains even if we consider more and more temporally fragile versions of the window breaking, and it remains even if we consider more and more intermediate times between  $t$  and the window breaking. So, condition (a) of CAUSATION is satisfied. Moreover, the window breaking is less secure and the

window's staying intact more secure in the closest-to-@-at-*t* world where Suzy does not throw her rock ( $w_8$ ) than in @. In  $w_8$ , the window would have stayed intact if just Billy had not been ready to intervene, whereas in @, the window would have stayed intact if Billy had not been ready to intervene and if Suzy had not thrown her rock. So, condition (b) is also satisfied.

Moreover, BLAMEWORTHINESS FOR similarly entails that Suzy is blameworthy for the window breaking rather than staying intact, given that we hold fixed that there was no gust of wind. We can use a slightly modified version of  $H_{U\text{-small}}$  that focuses on Suzy's and Billy's poor qualities of will instead of what they did. (i) That the window breaks is worse than its staying intact partly in virtue of the fact that it is worse that the elderly couple gets upset than it is that they do not. (Suzy has an the-elderly-couple-becoming-upset-related reason not to break the window.) Moreover, using a similar line of reasoning as we did when showing that Suzy's throw caused the window to break rather than staying intact, we can show that (ii) Suzy's poor quality of will towards the elderly couple was a cause of the window's breaking rather than staying intact. And, we might assume, her poor quality of will *non-deviantly* causes this outcome. The outcome is brought about roughly as Suzy planned, and there is no reason to think that causation went astray before she formed the intention to throw the rock.

To sum up, if we disregard the fact that Billy is lurking in the background with sinister intent, it seems that Suzy's throw caused the window breaking, and that she is blameworthy for breaking the window and upsetting the elderly couple. If we instead take seriously the fact that Billy is present, but hold fixed that there was no gust of wind (or any other intervening factor), it again seems that Suzy caused the window to break, and that she is blameworthy for doing so. If we conversely treat it as an open (and likely) possibility that Suzy's rock does not hit the window, and that the skilled thrower Billy would attempt to break the window in case Suzy did not attempt to do so, it seems that the window broke in spite of what Suzy did, and that she is not blameworthy for breaking the window. These are also the verdicts that CAUSATION and BLAMEWORTHINESS FOR gives. So, we have not found any counterexample to these principles. Rather, CAUSATION and BLAMEWORTHINESS FOR can explain the conflicting intuitions we might have about this case.

There are two questions remaining. First, which story is the more accurate one? As I already have argued, we should set aside the first story where we disregard the fact that Billy is present. Still, there are two plausible ways of understanding UNSKILLED SUZY. We could think that the window broke in spite of what Suzy did, or we might agree with the elderly couple that Suzy caused the window to break since there in fact was no gust of wind that made her throw go astray.

Notably, RELEVANT POSSIBILITIES FOR BLAME does not help us settle which story that is the more accurate one. This principle tells us to include, as a minimum, every combination of the actual poor quality of wills of the agents involved in the

situation, and the quality of will they were minimally required to have. Both  $H_{U\text{-large}}$  and  $H_{U\text{-small}}$  satisfy this requirement. Moreover, the principle WIDENING AND ADJUSTING tells us to look for more possibilities to include in our current possibility horizon if the outcome is unsatisfactorily explained (and if it is important to explain this outcome). However, the outcome does not seem unsatisfactorily explained in the smaller possibility horizon. We have found a cause of the outcome, and a verdict about who is blameworthy for it.

Ultimately, it is an open question which possibility horizon we should use when deciding whether Suzy's throw caused the window breaking and whether Suzy is blameworthy for breaking the window. Still, in cases where we are uncertain about which possibility horizon that is the correct one, we normally do best in considering the case closer, including more possibilities. Therefore, there is a reason to think that the larger possibility horizon is the more accurate one.

Second, there might be other ways of understanding UNSKILLED SUZY which I have not considered here. However, this case only constitutes a counterexample to CAUSATION and BLAMEWORTHINESS FOR if you think of it as a case where the window broke in spite of Suzy's throw (as captured by  $H_{U\text{-large}}$ ), but where Suzy's throw still is a cause of this outcome. Frankly, I do not see how this is possible. If the window broke in spite of what Suzy did, it seems completely wrong to say that what Suzy did was a cause of the window breaking.

Perhaps more helpfully, I can offer an error theory of sorts for why you might think that Suzy's throw caused the window breaking, even though the window breaking occurred in spite of her throwing the rock. Even though Suzy's throw made the window breaking less secure, it is still process-connected to the window breaking. It belongs to a set of events that guarantees that the window breaking will occur, and that does not guarantee this without Suzy's throw. And this picture remains even if we consider more and more temporally fragile versions of the window breaking, and it remains even if we consider more and more intermediate times between  $t$  and the window breaking. That is, there is a causal connection of sorts present, albeit not one that is sufficient for grounding blameworthiness for outcomes (as evinced by switching cases and other cases). This might explain why you might be tempted to think that Suzy's throw caused the window to break, even though it makes this outcome less secure.

## Conclusion

CAUSATION and BLAMEWORTHINESS FOR give the intuitively correct verdict in a wide range of cases, including non-threshold cases, climate change and PENNED-IN SHARKS. In addition, they can explain why our verdicts about causation and who is blameworthy for the outcome vary in these cases – the different verdicts stem from

different possibility horizons. This makes it less clear which causal verdict and which verdict about blameworthiness that are the more accurate ones. There are reasons, however, for thinking that the larger possibility horizon is the more accurate one in these cases. UNSKILLED SUZY might seem to be a counterexample to CAUSATION and BLAMEWORTHINESS FOR. However, if we specify how to understand this case, it turns out that these principles give intuitively correct verdicts.

# 14. Moral Entails Causal

In a few papers, and in her recent book (2016) *Causation and Free Will*, Carolina Sartorio argues that an agent can be morally responsible for an outcome without causing it. If she is correct, Touborg and I are mistaken in claiming that one is morally responsible for an outcome just in case one's substandard quality of will is a cause of this outcome. In this, the final chapter of my thesis, I argue that Sartorio's arguments are unsound. I start by considering the arguments she presents in (2004) "How to be Responsible for Something without Causing it", and then examine the considerations she adduces in (2015) "Resultant Luck and the Thirsty Traveler" and the already mentioned *Causation and Free Will*. The aim of this chapter is not just to defend BLAMEWORTHINESS FOR against Sartorio's pressing arguments, but also to open up an opportunity to discuss some interesting and challenging cases.

To give a hint of what is coming, in defending the idea that one is morally responsible for an outcome just in case one has caused it, I will argue that whether you cause an outcome might depend on whether you bring it about together with another agent, or together with some non-agential event, such as a mechanism or a natural phenomenon. I will also argue that CAUSATION explains our causal intuitions in the famous case of the thirsty traveller more satisfyingly than Sartorio's (2015, 2016) appeal to disjunctive facts as causes.

## Two Buttons

Sartorio (2004) seeks to disprove the following principle:

MORAL ENTAILS CAUSAL: If an agent is responsible for an outcome, it is in virtue of the fact that he caused it (some action or omission of his caused it).

(Sartorio 2004: 317)

Here, being responsible for some outcome is short for being *morally* responsible for it – that is, being blameworthy or praiseworthy for the outcome.

To disprove MORAL ENTAILS CAUSAL, Sartorio presents a counterexample, namely:



[TWO BUTTONS:] There was an accidental leak of a dangerous chemical at a high-risk chemical plant, which is on the verge of causing an explosion. The explosion will occur unless the room containing the chemical is immediately sealed. Suppose that sealing the room requires that two buttons – call them “A” and “B” – be depressed at the same time  $t$  (say, two seconds from now). You and I work at the plant, in different rooms, and we are in charge of accident prevention. Button A is in my room, and button B is in yours. We don’t have time to get in touch with each other to find out what the other is going to do; however, we are both aware of what we are supposed to do. As it turns out, each of us independently decides to keep reading his magazine instead of depressing his button. The explosion ensues.

(Sartorio 2004: 317)

Sartorio argues that the intuitive verdict in this case is that both you and I are morally responsible for the outcome.<sup>1</sup> I agree. Still, Sartorio argues, neither you nor I caused the outcome, and therefore this is a case of moral responsibility without causation, so TWO BUTTONS is a counterexample to MORAL ENTAILS CAUSAL. I disagree with this second claim. Intuitively, both you and I caused the explosion. So, it appears that TWO BUTTONS is not a counterexample to MORAL ENTAILS CAUSAL.

Sartorio gives a persuasive, but ultimately unsuccessful, argument for the claim that neither you nor I caused the outcome. I will now go through this argument and show where it goes wrong. This is the gist of Sartorio’s argument: there is a case that is identical to TWO BUTTONS in all causally relevant respects, and where it is obvious that I do not cause the explosion, and since I do not cause the explosion in this case, I do not cause the explosion in TWO BUTTONS.

The basic set-up in this other case is the same as in TWO BUTTONS: there are two buttons that need to be pressed in order to prevent an explosion. It goes as follows:

[TWO BUTTONS, ONE STUCK:] Again, button A is in my room, and I fail to depress it. This time, however, there is no one in the room containing button B; instead, a safety mechanism has been automatically set to depress B at  $t$ . When the time comes, however, B becomes stuck while being up. Just as in the original case, then, neither button is depressed and the explosion occurs.

(Sartorio 2004: 318)

Sartorio argues that I do not cause the explosion in this case. I agree. She also argues that the two cases are similar in all causally relevant respects. I disagree.

Sartorio gives two arguments for holding that the cases are identical in all causally relevant respects. The first is that “the only important difference between them is

---

<sup>1</sup> I adopt Sartorio’s language of “you and I”. Obviously, the “I” now refers to me, Mattias Gunnemyr, but nothing will turn on this in what follows.

that a person is in control of button B in one case but a mechanism is in control in the other” (323), and that such a difference cannot be causally relevant. Second, she argues that “the main existing theories of causation are likely to regard the two cases as causally on a par” (324), and that this indicates that the two cases are similar in all causally relevant respects. I am unconvinced by both arguments. Let us start by considering the second one.

### **So Much the Worse for the Main Theories of Causation**

Sartorio argues that both counterfactual and regularity theories of causation entail that the two cases are causally on a par. To be more precise, she argues that counterfactual theories entail that my failure to press the button is a cause of the explosion *in neither case*, and that regularity theories entail that my failure to press the button is a cause *in both cases*. So, whichever type of theory we apply we get the result that the two cases are causally on a par. This is true, but we should not take the fact that counterfactual theories entail that there is no causally relevant difference between these cases to indicate that there is no such difference. Counterfactual theories are notorious for giving unreliable verdicts in cases like these. Moreover, if regularity theories are correct in indicating that my failure to press the button is a cause of the explosion in both cases, we still have no counterexample to MORAL ENTAILS CAUSAL. Rather, we have an example where I seem to be blameworthy for the outcome *and* cause this outcome.

Let us consider these points more closely, starting with counterfactual theories. Sartorio understands these in the following way:

[V]ery roughly, and in its simplest version, a counterfactual theory deems something a cause when the effect counterfactually depends on it, i.e. had the cause not occurred, the effect wouldn't have occurred.

(Sartorio 2004: 324)

In other words, she takes counterfactual theories in their simplest form to be what I earlier have called SIMPLE.<sup>2</sup> She continues:

[... G]iven that the explosion would only have been prevented by depressing both buttons, and given that the other button wasn't depressed, the explosion would still have occurred if I had depressed A. So the explosion doesn't counterfactually depend on my failure to depress A, and thus a counterfactual theory would not count my failure to depress A as a cause of the explosion. Again, whether the other button

---

<sup>2</sup> This principle is introduced on p. 76.

wasn't depressed due to a human or a mechanical failure is simply irrelevant to the fact that the explosion doesn't counterfactually depend on my failure to depress A

(Sartorio 2004: 324)

Differently put, if we assume a counterfactual theory of causation, we find that the cases are indeed causally on a par. In both, the explosion is overdetermined: it would occur whether or not I press button A. In TWO BUTTONS, the explosion is overdetermined by my not pressing button A and your not pressing button B. In TWO BUTTONS, ONE STUCK, the explosion is overdetermined by my not pressing button A and the mechanism's failing to depress button B.

Now, as I have argued earlier, and as others have pointed out before me, counterfactual theories of causation are known to give counterintuitive verdicts in cases of overdetermination. We should not, therefore, rely on their verdicts in such cases. We should not refer to SIMPLE to show that I cause the explosion neither in TWO BUTTONS nor in TWO BUTTONS, ONE STUCK. Hence, should not use SIMPLE as a basis on which to draw the conclusion that there is no causally relevant difference between these two cases. Rather, we should postpone drawing any conclusions about causality in these cases until we have an account of causation that reliably gives correct verdicts in cases of overdetermination.

I now move on to the claim that regularity theories entail that the two cases are causally on a par. Sartorio depicts such theories in the following way:

Very roughly, and in its simplest version, a regularity theory deems something a cause when it is sufficient, in the circumstances, and given the laws, for the occurrence of the effect.

(Sartorio 2004: 324)

NESS fits this bill.

If we accept that omissions can be causes (something we should do, and which Sartorio does), regularity theories in their simplest form entail that my failure to press button A is a cause of the explosion in both cases. Sartorio is right in concluding this. So, regularity theories entail that the two cases are causally on a par. However, even if our best theory of causation turned out to be a regularity theory, TWO BUTTONS would not be a counterexample to MORAL ENTAILS CAUSAL. The theory would allow us to infer that the two cases are similar in all causally relevant respects, but it would licence that inference on the grounds that, in both cases, *I do cause the explosion*. And if it is true that I do cause the explosion in TWO BUTTONS, this is not a counterexample to MORAL ENTAILS CAUSAL. Rather, it is an example of a sequence in which I seem to be blameworthy for the outcome and also cause this outcome.

For these reasons, I do not find Sartorio's second argument convincing. My view is that turning to the main existing theories of causation does not help us to establish that TWO BUTTONS is a counterexample to MORAL ENTAILS CAUSAL. Let us instead consider her first argument.

### **It Might Matter Whether it is an Agent or a Mechanism**

Sartorio (2004) argues that "the only important difference between them [the two cases] is that a person is in control of button B in one case but a mechanism is in control in the other" (323). She contends that since this difference is causally irrelevant, the two cases are not different in any causally relevant sense.

There is, however, another relevant difference between the two cases – one that differs from the one that Sartorio mentions. As I understand TWO BUTTONS, ONE STUCK, there is a safety mechanism that will push button B. It does push the button, but it fails to depress it adequately because the button is stuck. Therefore, the mechanism's pushing button B does not help and the explosion ensues.

The fact that button B is stuck is causally relevant. If we add to TWO BUTTONS the detail that there is no possibility of avoiding the explosion because button B is stuck, we get a different causal verdict. For if there is no possibility of avoiding the explosion – if the explosion will ensue even if both you and I push our buttons – it seems incorrect to say that you and I caused the explosion.

For this reason, Sartorio is mistaken in thinking that the only relevant difference between TWO BUTTONS and TWO BUTTONS, ONE STUCK is that there is an agent operating button B in one case but a mechanism at work in the other. There is also a further relevant difference: that button B is stuck in one case but not in the other. In the light of this, we can adjust Sartorio's argument by slightly modifying TWO BUTTONS, ONE STUCK, as follows:

ONE AGENT, ONE MECHANISM: In order to avoid a disastrous explosion, both button A and button B need to be pressed before the critical period of time has passed. I am in control of button A and a safety mechanism has been automatically set to depress B before the critical period lapses. However, as the result of no human fault, the safety mechanism is broken. When the relevant time comes, I continue to read my magazine and the safety mechanism fails. The explosion ensues.

Here, with some hesitation, I am inclined to conclude that I did not cause the explosion. Moreover, the only difference between this case and the original TWO BUTTONS case is that here there is a mechanism in control of button B whereas in TWO BUTTONS there is an agent in control of button B. We might now reconstruct Sartorio's argument using this example. We can argue that in ONE AGENT, ONE MECHANISM it is obvious that I do not cause the explosion. We can insist that the

only difference between ONE AGENT, ONE MECHANISM and TWO BUTTONS is that there is a person in control of button B in one case but a mechanism in the other. And we can argue that this difference cannot make a causally relevant difference, and that therefore I do not cause the explosion in TWO BUTTONS.

I think this attempt to improve the argument is unsuccessful. The fact that a person is in control of button B in one case but a mechanism in the other might well be causally relevant. How could this be so? Well, to begin with we can note that, apparently, it *is* relevant. Remember that our intuitive verdict about TWO BUTTONS was that my failure to press button A is a cause of the explosion (as is your failure to press button B). However, in ONE AGENT, ONE MECHANISM the intuitive verdict is that my failure to press button A is *not* a cause of the explosion. So, the evidence we have indicates that it does matter, when we are arriving at verdicts about causation, whether a person or a mechanism is in charge of button B.<sup>3</sup>

## Agents, Mechanisms and Relevant Possibilities

CAUSATION can explain our intuitions about causation in these cases.

CAUSATION: Suppose that C occurs at  $t$  and E occurs later, that C\* is a merely possible event that is incompatible with C, and that E\* likewise is a merely possible event that is incompatible with E.

Then C rather than C\* is a cause of E rather than E\* within possibility horizon H just in case

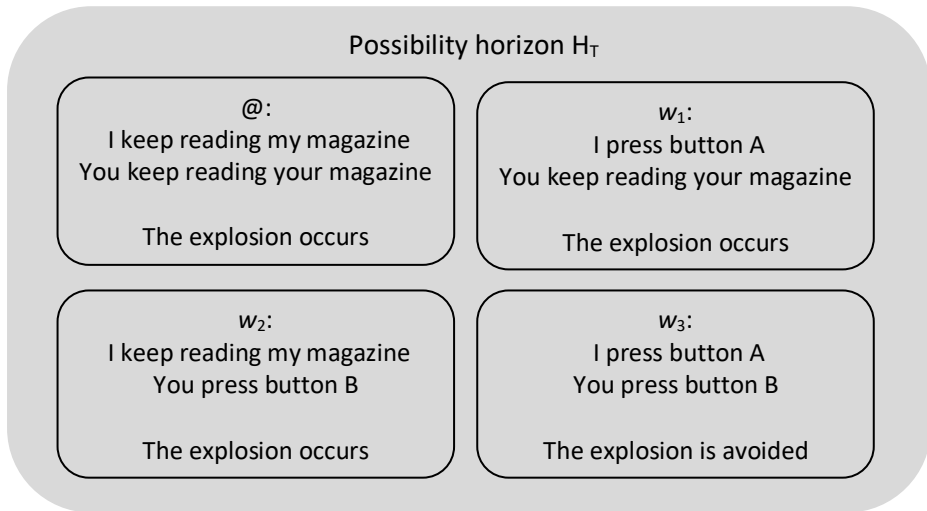
- (a) C is process-connected to E,
- (b) there is at least one world in H where C occurs at  $t$ , and at least one world in H where C\* occurs at  $t$ , and E is more secure, and E\* is less secure, in the closest-to-@-at- $t$  world(s) in H where C occurs at  $t$  than they are in the closest-to-@-at- $t$  world(s) in H where C\* occurs at  $t$ .<sup>4</sup>

To apply CAUSATION to a particular case, we first have to decide upon the relevant possibility horizon. In TWO BUTTONS there are four relevant possibilities, here depicted as four relevant possible worlds at the relevant time  $t$  (which could be any time within the critical time period when the two buttons must be pressed in order for the explosion not to occur).

---

<sup>3</sup> Sartorio illustrates the idea it does not make a difference, for our causal verdicts, whether the other potential cause was a sentient being or a purely physical mechanism with other cases. I have here decided not to discuss these other cases, as I believe that doing so would not add any additional considerations to the discussion.

<sup>4</sup> This is the updated version of CAUSATION presented towards the end of Chapter 12.



For the sake of clarity, I should say that I use the following translations in the argument that ensues.

- C = I keep reading the magazine
- C\* = I press button A
- E = The explosion occurs
- E\* = The explosion is avoided

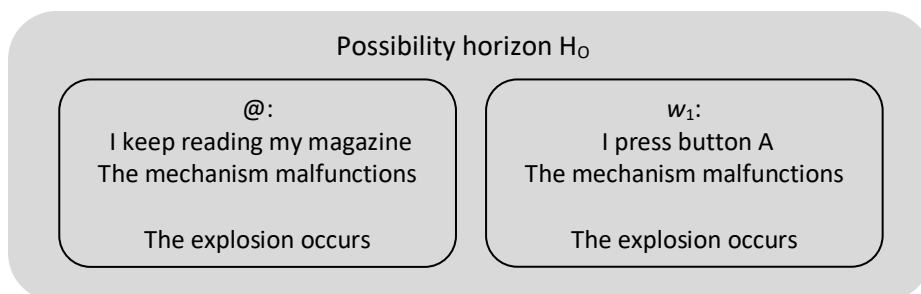
According to CAUSATION, the fact that I kept reading my magazine rather than pressed button A at time  $t$  was a cause of the explosion's occurring rather than not doing so. First, my reading of the magazine is process-connected to the explosion. It belongs to a set of concurrent events that guarantees, given the laws of nature, that the explosion will occur, and this set no longer guarantees, given the laws of nature, that the explosion will occur if I press button A. Moreover, this picture remains intact even if we consider more and more temporally fragile versions of the explosion, and it does so even if we consider more and more intermediate times between  $t$  and the explosion. So, condition (a) of CAUSATION is satisfied.

Second, (b) there is at least one world in  $H_T$  where I keep reading my magazine, and at least one world where I press button A. Moreover, the occurrence of the explosion is more secure, and the avoidance of the explosion less secure, in the closest-to-@-at- $t$  world in  $H_T$  where I keep reading my magazine (which is @) than they are in the closest-to-@-at- $t$  world in  $H_T$  where I press button A (which is  $w_1$ ). If I had pressed the button (as in  $w_1$ ), the only thing that would have needed to change in order for the explosion not to occur is that you press your button rather than continue reading your magazine. By contrast, in the actual world, one more thing would need

to change in order for the explosion not to occur: in addition to your pressing your button, I would need to press mine. So, condition (b) is also satisfied.

This means CAUSATION entails that my failure to press button A at time  $t$  is a cause of the explosion. It yields the same verdict as a simple regularity account of causation, but a verdict that differs from that given by a simple counterfactual account. Importantly, the verdict given by CAUSATION matches our intuitive thinking about the case.

Let us now see what CAUSATION says about ONE AGENT, ONE MECHANISM. First, we have to settle on the appropriate possibility horizon. The possibility horizon differs in one crucial respect between the two cases. Since the safety mechanism is broken in ONE AGENT, ONE MECHANISM, we can safely assume that there is no possibility that the mechanism will press button B. Given this, we get the following, much-reduced possibility horizon.



We can now see that in ONE AGENT, ONE MECHANISM, CAUSATION entails that I did not cause the outcome. Although there is a process-connection between my reading the magazine and the explosion, my reading the magazine rather than pressing button A does not increase the security of the explosion. The explosion is just as secure, within  $H_0$ , whether I read the magazine or press button A – namely, infinitely secure. Since there is no possible world within  $H_0$  where the explosion does not occur, the distance between the actual world where I read the magazine and the closest possible world where the explosion does not occur is infinite, and for the same reason the distance between the closest possible world where I press the button and the closest possible world where the explosion does not occur is also infinite. In short, although (a) is satisfied, (b) is not. Fundamentally, the causally relevant difference between TWO BUTTONS and ONE AGENT, ONE MECHANISM is that, in the former, there is a possibility that the explosion does not occur, whereas in the latter there is no such possibility. To the extent that we can rely on CAUSATION giving accurate decisions on causation, there is this causally relevant difference

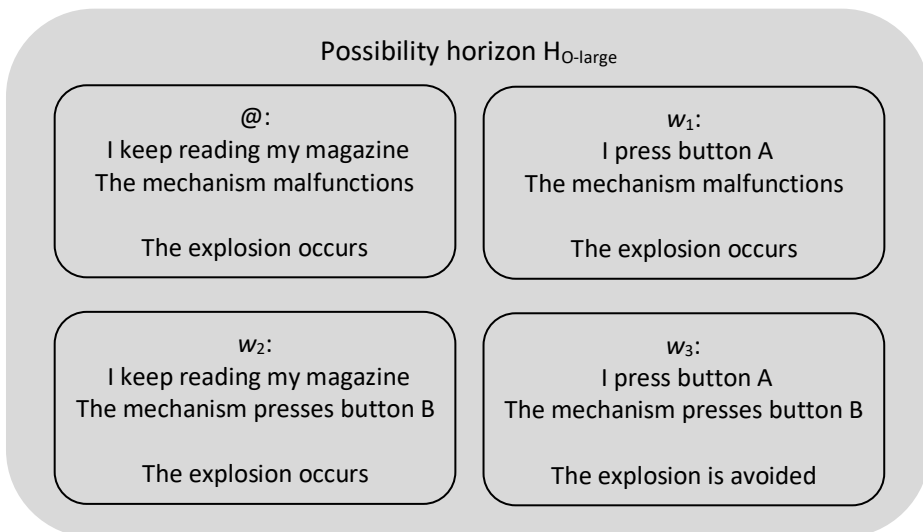
between the two cases, which means that Sartorio’s argument to the contrary is unsuccessful.

### Relevant Possibilities for Blame

We could bring our enquiries into TWO BUTTONS to an end here. However, there is a further question it would be helpful to sort out. The question is this: Why do we treat your pressing the button as a possibility in TWO BUTTONS, yet decline to treat the safety mechanism’s not malfunctioning as a possibility in ONE AGENT, ONE MECHANISM?

Remember that I said that I only hesitantly think that I am not responsible for the explosion in ONE AGENT, ONE MECHANISM. There was a reason for this. There is an alternative way to understand this case. If we think that the fact that the safety mechanism is broken excludes the possibility that the explosion will not occur, we get the much-reduced possibility horizon  $H_O$ , which means we can then conclude that I did not cause the explosion in this case. If, however, we entertain the possibility that the safety mechanism is not broken, the case appears in a different light. If it is possible for the safety mechanism to function properly, the intuitive verdict will be that I did have a reason to push the button, that I am morally responsible for the ensuing explosion, and that I caused the explosion by not pressing the button. These verdicts remain sound even if the mechanism turns out not to function on this particular occasion.

CAUSATION can explain these conflicting intuitions. If we assume it is possible that the mechanism will work – i.e. will not malfunction – we get an expanded possibility horizon.





This possibility horizon closely resembles that in TWO BUTTONS. We also get the same causal verdict when applying CAUSATION to it as the one we obtained vis-à-vis TWO BUTTONS. In other words, with this possibility horizon, I *do* cause the explosion in ONE AGENT, ONE MECHANISM. Thus, (a) my reading the magazine is process-connected to the explosion (just as it is in TWO BUTTONS). (b) The explosion is now more secure in @ and the avoidance of the explosion is less secure in @ than they are in the closest-to-@-at-*t* world where I press button A (i.e.  $w_1$ ). If I had pressed A, the only thing that would have needed to change for the explosion not to occur is your pressing B instead of reading your magazine; whereas in @ you would have to press B and I would have to press A in order for the explosion not to occur.

So, which possibility horizon is the correct one? Did I cause the explosion in ONE AGENT, ONE MECHANISM? The perhaps unsatisfying answer is that I caused the explosion within the larger possibility horizon that includes the possibility that the safety mechanism functions properly, and that I did not cause the explosion within the smaller possibility horizon that treats it as a background condition that the safety mechanism is broken. Metaphysically speaking, that is all there is to say. The possibility horizon is one of the causal relata in CAUSATION. Depending on our interests and perspective, it may or may not be appropriate to include the possibility that the mechanism might have functioned properly. If we are button-technicians working at the chemical plant, we should certainly entertain the possibility that the button did not malfunction. If, instead, we are asking who is to blame for the explosion, this matter is less clear.

In “You Just Didn’t Care Enough”, Touborg and I suggest that when we are assessing blame, we should at least include the involved agents’ actual quality of will as well as the quality of will that they should have in the relevant possibility horizon. (I repeat the relevant principle, consisting of four claims, for ease of reference.)

RELEVANT POSSIBILITIES FOR BLAME: To determine, for the purpose of attributing blame, whether your poor quality of will at time *t* (in relation to Y versus Y\*) is a cause of a later event X rather than X\*, it is a relevant possibility that you could instead have had the minimally required quality of will at *t* (in relation to Y versus Y\*). Similarly, it is a relevant possibility that anyone else involved in the situation who had a poor quality of will at time *t* could have had the minimally required quality of will at time *t*. Every combination of these possibilities is relevant. Other possibilities may or may not be relevant as well.<sup>5</sup>

This principle entails that we have to include the possibility that you and I both press our buttons in TWO BUTTONS (which we did). In this case, there is a clear answer to the question of whether I caused the explosion, and whether I am blameworthy for

---

<sup>5</sup> We suggest a similar principle in “Reasons for Action”.

it. I did and I am. RELEVANT POSSIBILITIES FOR BLAME also entails that we do *not have to* include the possibility that the mechanism does not malfunction in ONE AGENT, ONE MECHANISM. On the other hand, it does not say that we are not permitted to include that possibility. So, the questions whether I caused the explosion, and whether I am blameworthy for it, remain open in ONE AGENT, ONE MECHANISM.

In the end, I think the larger possibility horizon is the more correct one. As Touborg and I argue in “You Just Didn’t Care Enough”, if it is important to decide what caused a certain outcome, and if the causes of outcome are unsatisfactorily explained, we have reasons to widen and adjust our possibility horizon. In this case, this suggests that if it is important to decide what caused the outcome (as it seems to be), and if the causes of the outcome are unsatisfactorily explained (as seems to be the case given that we cannot decide which possibility horizon to use), we should opt for the larger possibility horizon.

Returning to Sartorio’s argument, I wish to point out that whether we use the smaller or larger possibility horizon in ONE AGENT, ONE MECHANISM, we will not arrive at a counterargument to MORAL ENTAILS CAUSAL. For if we adopt the smaller possibility horizon, there is a causally relevant difference between this case and TWO BUTTONS: in one case, but not the other, there is a possibility that the explosion will not occur. Therefore, Sartorio’s claim that the two cases are causally on a par is untrue. If, on the other hand, we apply the larger possibility horizon in ONE AGENT, ONE MECHANISM, Sartorio is correct that there is no causally relevant difference between the two cases. However, it is no longer true that I did not cause the explosion in ONE AGENT, ONE MECHANISM, and therefore we cannot argue that I did not cause the explosion in TWO BUTTONS on the basis that the two cases are causally on a par. Rather, we must conclude that I caused the explosion in both cases. Again, then, we fail to identify a counterexample to MORAL ENTAILS CAUSAL.

## The Thirsty Traveller

Sartorio (2015, 2016) recasts the argument against MORAL ENTAILS CAUSAL with an example that has been extensively discussed in the literature on causation in the law – namely, the case of the thirsty traveller. The example comes in many variants.<sup>6</sup> Sartorio (2016) presents it in the following way:

---

<sup>6</sup> The case was first introduced by McLaughlin (1925-26). For discussion, see e.g. Hart and Honoré (1985), Gavison et al. (1980), Mackie (1974), Wright (1985, 2013), Kvat (2002), Stapleton (2008), M. S. Moore (2009), Sartorio (2015, 2016), Talbert (2015) and Bernstein (2019).

[THE THIRSTY TRAVELLER: A] ... man fills his canteen with water before taking a trip into the desert. The man has two enemies who want him dead. The first enemy secretly drains the water out of the canteen and replaces the water with sand. A bit later, a second enemy, unaware of what the first enemy has done, steals the canteen from the man. The man then dies of thirst.

(Sartorio 2016: 27)

Building on earlier work (Sartorio 2006, 2015), Sartorio (2016) argues that neither enemy caused the traveller's death. She reasons that "stealing a canteen filled with sand cannot causally result in a death from thirst. The same goes for draining water out of a canteen that will be miles away from the man when he needs it" (27). Still, she continues, there is a puzzle here since the traveller clearly died as a result of what his enemies did. On her preferred solution to the puzzle, neither enemy caused the traveller's death although both were involved in the causal history of it. Building on this idea, she argues that although neither enemy (call them A and B) caused the death of the traveller, at least one of them is morally responsible for the death of the traveller, and is so in virtue of being part of the causal history of the outcome. If Sartorio is right, THE THIRSTY TRAVELLER is a counterexample to MORAL ENTAILS CAUSAL. At least one enemy is morally responsible for the traveller's death without having caused it.

I think this is the wrong way to think about the case. At the general level, THE THIRSTY TRAVELLER is a puzzle about both causation and blameworthiness. When we consider A and B separately, it seems that neither of them caused the death of the traveller and that neither of them is blameworthy for it. However, when we consider them as a group, it seems that the traveller died as a result of what they did and that at least one of them is blameworthy for the death of the traveller. This difference is particularly salient when we consider the possibility that neither of them had acted as they did. The fact that our intuitions about causation and blameworthiness go one way when we consider A and B separately, and another when we consider them as a group, suggests that there is a common solution to both puzzles, not one that only resolves that causal puzzle but leaves the moral one unsolved.

Still, Sartorio makes a strong case for thinking that in THE THIRSTY TRAVELLER someone is blameworthy for an outcome without having caused it. Here, I will present her arguments in detail, and explain why I think they fail. As an important part of my argument, I will try to show that CAUSATION and BLAMEWORTHINESS FOR can explain the intuitions we have about THE THIRSTY TRAVELLER.

## Sartorio's arguments

Sartorio's (2015) main goal is to show that there is a previously unexplored form of resultant moral luck that has little to nothing to do with whether you cause an outcome. In the course of arguing for this conclusion, she also argues that you might be blameworthy for an outcome without causing it. Here, I will concentrate on the arguments concerning the second issue.

Sartorio's (2015) strategy for establishing that neither A nor B caused the death of the traveller resembles the one she uses to show that neither you nor I cause the explosion in TWO BUTTONS. She suggests that since our intuitions about causation in THE THIRSTY TRAVELLER are uncertain, we should instead reflect on cases that are identical to THE THIRSTY TRAVELLER in all causally relevant respects, but where our intuitions are clearer. To decide whether A caused the death of the traveller, she compares the original case with the following one (here presented in slightly amended form):<sup>7</sup>

THE THIRSTY TRAVELLER WITH A FRIEND: Everything is as in THE THIRSTY TRAVELLER, except that A is not the traveller's enemy but his best friend. A knows about B's plan to steal the canteen, and tries to persuade his friend not to follow through on his plan to go into the desert. The traveller, however, refuses to believe that someone is planning to kill him, and remains determined to stick with his plan. Sadly realising that he cannot prevent the death of his beloved friend, A replaces the water in the canteen with sand so that B will also die of thirst (A knows that B will be counting on the water in the stolen canteen to survive on his way back from the desert). As predicted, B steals the canteen, the traveller dies from thirst, and so does B.

In this modified case, Sartorio argues, it is clear that while A caused B's death, he did not cause the death of the traveller. And, since there is no causally relevant difference between this case and the original case of THE THIRSTY TRAVELLER (because differences in beliefs, intentions, and so on, are causally irrelevant, she insists), A did not cause the death of the traveller in the latter. In a similar fashion, appealing to another variation of THE THIRSTY TRAVELLER in which B is the traveller's best friend but A is not, Sartorio argues that B did not cause the death of the traveller in the original case. Call this *the comparison argument*.

Bear in mind Sartorio's (2016) point that draining water out of a canteen that will be miles away from the traveller when he needs it cannot causally result in the traveller's dying from thirst, and that stealing a canteen filled with sand likewise cannot causally result in the traveller's dying from thirst. Sartorio (2006) once

---

<sup>7</sup> In the version of THE THIRSTY TRAVELLER Sartorio (2015) considers, A replaces the water with salt instead of sand.

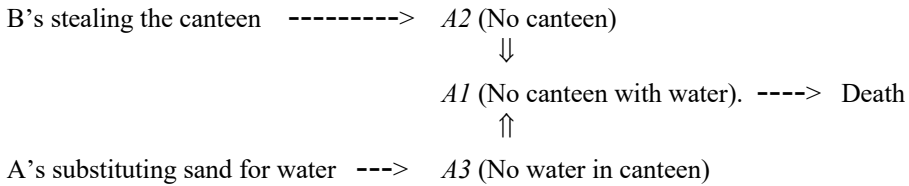
elaborated these ideas, albeit in relation to another (but similar) example, effectively arguing that A's draining the water out of the canteen and replacing it with sand *would have been* a cause of the death of the traveller if B had not stolen the canteen, but that it fails to cause the death because, as things turned out, B did steal the canteen.

Sartorio makes a similar point about B: B's stealing a canteen filled with sand cannot causally result in the traveller's dying from thirst. She argues that although B's stealing the canteen *would have been* a cause of the death of the traveller if A had not replaced the water with sand, it fails to cause the death because, as things turned out, A did replace the water with sand.

As Sartorio sees it, while A and B are both in some way involved in the causal history of the traveller's death, what each does merely amounts to making a difference to *how* the death of the traveller comes about, not *whether* he will die. In effect, A's replacing the water with sand works as a switch. It changes the causal history of the death of the traveller, but it makes no difference to whether the traveller will die, so is not a cause of the traveller's death. In this respect, it is like the flipping of the switch in THE ENGINEER (introduced on p. 27). Likewise, B's stealing of the canteen makes a difference to the causal history of the death of the traveller, but not to whether the traveller will die, and therefore it is not a cause of the traveller's death. Call this the *mere switches argument*.

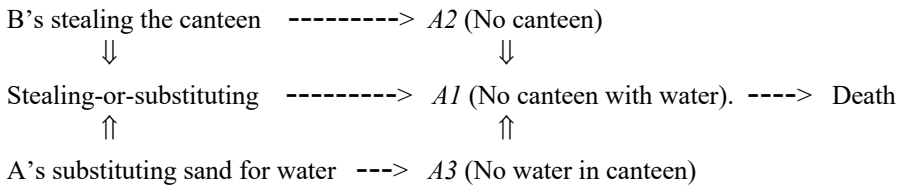
If the comparison argument and the mere switches argument are sound – indeed, if at least one of them is successful – it follows that neither A nor B caused the traveller's death. This might seem strange: it certainly seems that the traveller died because of what A and B did. Sartorio (2015) acknowledges this point, but she argues that it might be true that the cause of the traveller's death has something to do with what A and B did even if neither the doings of A nor the doings of B caused this death. Building on earlier work (Sartorio 2006), she suggests that we should think about the causal structure in this case in the following way.

The traveller's death was caused by the absence of *a canteen with water* at the relevant place and time. Call this absence *A1*. While B's stealing the canteen did not cause the traveller's death, it caused the absence of *a canteen* at the relevant place and time (call this *A2*), an absence which in turn logically implies that there was an absence of a canteen with water at the relevant place and time (that is, *A2* implies *A1*). Further, while A's replacing the water with sand did not cause the traveller's death, it caused the absence *of water* in the canteen at the relevant time (call this *A3*), an absence which also logically implies that there is an absence of a *canteen with water* at the relevant place and time (that is, *A3* also implies *A1*). These thoughts are illustrated below. The dotted arrows are causal relations and the double arrows are logical relations.



(adapted from Sartorio 2015)

Finally, while neither the stealing nor the substituting caused the absence of a canteen with water at the relevant place and time (A1), this is not an uncaused event. Instead, a disjunctive fact – the stealing-or-substituting – caused A1, and by extension the death of the traveller. Here, the disjunctive fact, the stealing-or-substituting, is “the fact that would have obtained just in case either [A] had substituted the [sand] for water when he did, or [B] had stolen the canteen when he did, or both” (Sartorio 2015: 13). If we add this to our illustration, we get the following picture:



(adapted from Sartorio 2015)

If this is an accurate description of the causal structure of the case, we can conclude that neither A nor B caused the traveller's death, and still allow that the traveller's death had something to do with what A and B did. There is no causal relation between B's stealing the canteen and the death of the traveller – neither a direct one nor an indirect one via some intermediate event. Likewise there is no causal relation between A's substituting sand for water and the traveller's death. But there is a causal relation between what A and B did – the stealing-or-substituting – and the death of the traveller. Call this the *explanation of the twin intuitions*. With this, we have an account of the causal structure of THE THIRSTY TRAVELLER.

Next, Sartorio (2015) argues that at least one of A and B is morally responsible for the traveller's death. If we have described the causal structure of the case accurately, the following answer to the question which of A and B this is suggests itself: “whoever is [morally] responsible for the cause of the death (the disjunctive fact, the stealing-or-substituting) is also [morally] responsible for the death” (16). Still,

it is not an easy task to identify who this is – who is morally responsible for the disjunctive fact. Sartorio considers several ways of deciding this in a principled way, and rejects most of them. For instance, she considers the proposal that anyone who is morally responsible for P is also morally responsible for Q if P logically implies Q, and the proposal that anyone who is morally responsible for P is also morally responsible for the disjunctive fact P-or-Q for any Q. She rejects both. (Why she rejects them is of less importance here.)

In the end, Sartorio says that we have reasons to believe – at least, tentatively – that, in cases like THE THIRSTY TRAVELLER, where what each agent does logically guarantees the occurrence of the cause of the outcome, and where this is the whole extent of their causal involvement with regard to the outcome, “whoever guaranteed the occurrence of the cause first bears all the responsibility” (21). That is, whoever is the *first* to perform an act that logically implies the disjunctive fact is morally responsible for the outcome.<sup>8</sup> When this principle is applied, it transpires that A, but not B, is morally responsible for the death of the traveller. A’s replacing the water with sand logically guarantees that the disjunctive fact stealing-or-substituting obtains, and it does so before B’s stealing of the canteen does the same. Therefore, A is morally responsible for the occurrence of the cause of the death of the traveller, which is the stealing-or-substituting, and by extension, A is also morally responsible for the death of the traveller. Call this the *explanation of moral responsibility*.

We now have two results that can be combined to refute MORAL ENTAILS CAUSAL. First, the comparison argument and the mere switches argument show that neither A nor B caused the death of the traveller. Second, the explanation of moral responsibility indicates that A is morally responsible for the outcome. If these conclusions are correct, A is morally responsible for the death of the traveller without having caused it, and thus THE THIRSTY TRAVELLER is a counterexample to MORAL ENTAILS CAUSAL. On top of this, we have an explanation of why it seems that neither A nor B caused the death of the outcome when we consider A and B separately, but that it also seems that the traveller died as a result of what A and B did if we consider them as a group. That is, we have an explanation for our twin intuitions about the case.

Importantly, it does not matter for Sartorio’s argument that it is A (and not B) that is morally responsible for the death of the traveller. The important point is that at least one of them is. This is enough to establish that in THE THIRSTY TRAVELLER *someone* is blameworthy for the death of the traveller without causing that death, which is enough to refute MORAL ENTAILS CAUSAL.

---

<sup>8</sup> I take it that it is presupposed that the agent also is morally responsible for acting in the relevant way.

## Disjunctive Causes

Do we have reasons to believe that disjunctive facts can be causes? Sartorio (2006) argues that we do (but again in relation to the other, similar, examples). One consideration is that a straightforward counterfactual account of causation like SIMPLE entails that, while neither A's substituting nor B's stealing caused the death of the traveller, the disjunctive fact – the stealing-or-substituting – did. While the death of the traveller would have occurred whether A had substituted sand for water or not, and whether or not B had stolen the canteen, it is also true that if the disjunctive fact had not obtained – that is, if neither the substituting nor the stealing had occurred – the traveller would not have died.

Sartorio (2006) is not the first to suggest that disjunctive facts can be causes. Alan Penczek (1997) argues that an outcome has a disjunctive cause in any case where there is more than one route to it. Sartorio (2006) finds this hard to believe. She points out that pre-emption cases are clear-cut counterexamples to this idea. Consider BOTTLE SHATTERING again,<sup>9</sup> where Suzy and Billy throw rocks at a bottle and Suzy hits the bottle first, shattering it. In this case there is more than one potential route to the bottle shattering. Billy's throw could have caused it. Equally, Suzy's throw could have done so. However, there is no need to posit a disjunctive cause. Indeed, it seems odd to say that the disjunctive fact Suzy-throws-or-Billy-throws caused the bottle shattering. Rather, Susy's throw caused the bottle to shatter, and Billy's throw did not. Unlike Penczek, Sartorio (2006) concludes that we only have reasons to posit disjunctive facts as causes in cases with a switching structure, like THE THIRSTY TRAVELLER.<sup>10</sup> As she puts it: “the suggestion is that the argument for disjunctive causes only generalizes to cases with the switching structure” (532). By restricting her argument in this way, she can avoid the troublesome conclusion that disjunctive facts are causes also in cases of pre-emption.

## Why Sartorio's Argument Fails

I deny that THE THIRSTY TRAVELLER is a counterexample to MORAL ENTAILS CAUSAL. To begin with, Sartorio's arguments for thinking that neither A nor B caused the death of the traveller are less than compelling. Consider first the comparison argument. In setting out this argument, Sartorio asks us to consider THE THIRSTY TRAVELLER WITH A FRIEND, where A is the traveller's best friend but replaces the water with sand anyway since he cannot persuade his friend to abstain from travelling into the desert, and since he wants B to die. This case, Sartorio

---

<sup>9</sup> This case was introduced on p. 81.

<sup>10</sup> Again, Sartorio (2006) does not say anything about THE THIRSTY TRAVELLER. Instead, she considers a case with a similar causal structure. Still, judging by her comments in later works (Sartorio 2015, 2016), it seems safe to conclude that she would accept that THE THIRSTY TRAVELLER exhibits the switching structure.



argues, is identical to THE THIRSTY TRAVELLER in all causally relevant respects. Further, because it is obvious that what A does in this, the modified scenario, is not a cause of the traveller's death, we are entitled to conclude that what A did in the original scenario was not a cause of the traveller's death.

I agree that A does not cause the traveller's death in THE THIRSTY TRAVELLER WITH A FRIEND. However, there is a causally relevant difference between the two cases. In the modified case, unlike the original one, *B does cause* the traveller's death. In the original case, the two enemies act independently of each other. In the modified case, by contrast, A replaces the water with sand because B plans to steal the canteen. We can see this difference by applying SIMPLE. In the modified case, whether the traveller dies depends on whether B plans to steal the canteen. Had B not planned to do so, A would not have replaced the water with sand, and consequently the traveller would not have died. In the original case, however, the traveller's death is not dependent on whether B is planning to steal the canteen. As a result of A's actions, the traveller will die whether or not B steals the canteen. If this is right – if there is a causally relevant difference between the two cases – we cannot infer, as Sartorio does, from the fact that A did not cause the traveller's death in the modified case, that A does not cause the traveller's death in the original one. Similar remarks could be made about Sartorio's parallel argument for the claim that B does not cause the traveller's death in the original case (an argument I have not introduced here, but only alluded to).

Consider next the mere switches argument. This argument essentially categorises the actions of A and B as mere switches and then infers that, since mere switches fail to make a difference to the outcome, A and B are not causes. I think Sartorio is right that mere switches are not causes, but I do not accept that what A does is a mere switch. A's replacing the water with sand does make a difference to the outcome (and not just the way it comes about): it makes a modal difference to the death of the traveller. While it is true that the traveller would have died whether or not A replaced the water with sand, it is also true that, in the nearby possible world where B does not steal the canteen, A's replacing the water with sand *does* make a difference to whether the traveller dies or not. As Sartorio notes, A's replacing the water with sand *would be* a cause if B had not stolen the canteen. In the framework within which I am working this is telling. In effect, Sartorio is saying that A's replacing the water with sand makes the death of the traveller *more secure*.

For comparison, consider once again THE ENGINEER. An engineer flips a switch, making a train travel down the right-hand track instead of the left-hand track, but the tracks later reconverge so the train arrives at its destination just the same. In this case the engineer's flipping of the switch is a mere switch.<sup>11</sup> It makes no modal

---

<sup>11</sup> Perhaps needless to say, the fact that the engineer flips a *switch* is coincidental. She might as well have pushed a button or pulled a lever.

difference to whether the train arrives at its destination. So, assuming the description of the case includes all relevant facts, it is untrue to say that the engineer's flipping of the switch would have caused the train's arrival at its destination if certain other events had turned out differently.

So, assuming that modal differences might matter (as I think we should), A's replacing the water in the canteen with sand is not a mere switch. Therefore, even if we accept that mere switches are not causes (as, again, I think we should), we cannot infer from this that A's replacing the water with sand is not a cause. What A did was not a mere switch. His replacing the water with sand made a difference to the outcome – the death of the traveller. Something similar can be said about B's stealing the canteen. While it is true that the traveller would have died whether or not B stole the canteen, it is also true that in the nearby possible world where A does not replace the water with sand, B's stealing the canteen makes a difference to the outcome – the survival of the traveller.

Since neither the comparison argument nor the mere switches argument is valid, we cannot conclude with Sartorio that neither A nor B caused the death of the traveller. There certainly seems to be something to the idea that neither A nor B caused the death of the traveller. However, we should not accept this idea unconditionally. Rather, we should take our initial causal intuitions seriously. We should take seriously the fact that, when we consider what A and B did individually, it seems that neither A nor B caused the death of the traveller, even though it also seems that what A and B did is in some way causally related to the traveller's death. After all, if neither A nor B had acted as they did, the traveller would have survived. These are the intuitions an accurate account of causation should be able to explain.

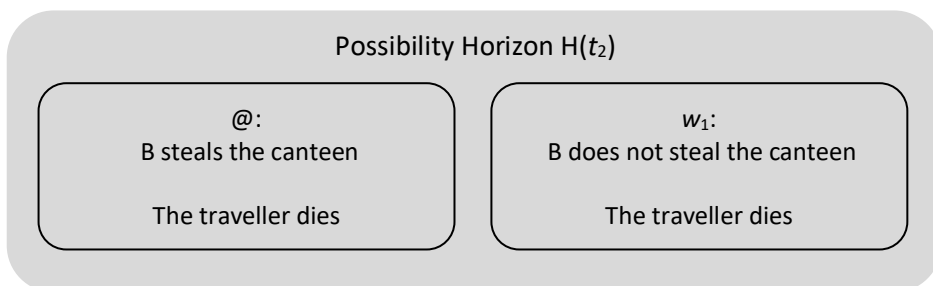
Of course, Sartorio is able to explain the twin intuitions. On her analysis, neither A nor B caused the death of the traveller, but what each of them did logically guaranteed the disjunctive fact that caused the traveller's death. Even though the comparison argument and the mere switches argument do not survive serious scrutiny, Sartorio's analysis might still provide the best explanation of our causal intuitions in THE THIRSTY TRAVELLER. This raises the question whether Sartorio's analysis offers the best explanation of these causal intuitions. I think not. The best way to show this would be to show that the Sartorio analysis generates counterexamples. However, because she restricts her analysis to apply only to cases with a structure like that in THE THIRSTY TRAVELLER, and because her analysis explains our intuitions well enough in such cases, it is hard (if not impossible) to show that it generates counterexamples.

Instead of arguing by way of counterexamples, I will therefore adopt a different line of attack. Usually, in any given case, more than one account of causation will give intuitively correct verdicts about what causes what. To take an example close to hand, both SIMPLE and CAUSATION deliver the right verdict about what causes what in a switching case like THE ENGINEER (that the engineer's flipping the switch is not

a cause of the train’s arrival at the station). However, when we are looking for the best explanation of our causal intuitions, we are not just looking for an account issuing in the right verdict about the case at hand. Rather, we want an explanation that delivers the right verdict in this case *and* across a wide range of other kinds of case. So, if CAUSATION can explain our causal intuitions in THE THIRSTY TRAVELLER and in a wide range of other cases, it provides a better explanation of our causal intuitions in THE THIRSTY TRAVELLER than Sartorio’s account does. As I have already argued, CAUSATION does give correct verdicts about causation across a wide range of cases. I will now try to show that it can also explain our twin intuitions about causation in THE THIRSTY TRAVELLER. If it can, we have strong reasons to believe that CAUSATION gives a better explanation for the causal intuitions in this case than Sartorio’s account does.

### Who Caused the Death of the Thirsty Traveller?

Consider first B. At the time he steals the canteen,  $t_2$ , there is no possibility that the traveller will survive. The possibility horizon contains only two possibilities: either he steals the canteen and the traveller dies from thirst, or he does not steal the canteen and the traveller dies from thirst.

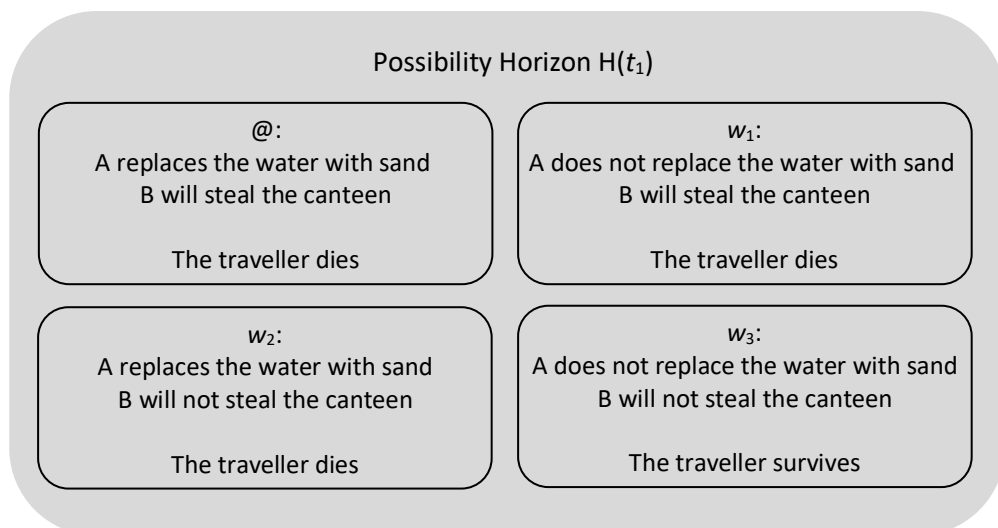


Given this possibility horizon, THE THIRSTY TRAVELLER is a switching case (just as THE ENGINEER is). B has control over whether the thirsty traveller will find a canteen filled with sand or no canteen at all, but that is all – he has no control over whether the traveller will die from thirst. In such cases, we usually judge that the agent does not cause the outcome. In this respect, Sartorio is correct when she says that THE THIRSTY TRAVELLER has a switching structure, and that “stealing a canteen filled with sand cannot causally result in a death from thirst” (2016: 27).

CAUSATION also yields the result that B did not cause the death of the traveller, given  $H(t_2)$ . B’s stealing the canteen does not increase the security of the death of

the traveller. His death is just as secure whether B steals the canteen or not. So, condition (b) of CAUSATION is not satisfied.

Now consider A. CAUSATION entails that A did not cause the death of the traveller, albeit for another reason. At  $t_1$ , the time at which A replaces the water with sand, there is still a possibility that the traveller will survive. The traveller will survive if A does not replace the water with sand and if B does not steal the canteen at  $t_2$ . We must therefore consider a larger possibility horizon for A:



Given this possibility horizon, we can see that A's replacing the water with sand does increase the security of the death of the traveller. Had he not replaced the water with sand, only one thing would need to change in order for the traveller to survive. In that case, the traveller will survive if B does not steal the canteen. For the same reason, replacing the water with sand decreases the security of the traveller's survival. The traveller's survival becomes an event that is even further from happening when A replaces the water with sand. So, condition (b) of CAUSATION is satisfied.

However, A's substituting sand for water is not process-connected to the death of the traveller. Both Richard Wright (2013) and John L. Mackie (1974) make a similar point, but in relation to the NESS condition and the INUS condition of causation, respectively. Wright considers a variant of THE THIRSTY TRAVELLER in which A does not replace the water with sand but adds a fatal dose of undetectable poison to

it.<sup>12</sup> Considering this modified case, he argues that A's adding poison to the water in the canteen does not satisfy the NESS condition of causation. In order for an event C to be a cause of E, according to NESS, both C and E must be instantiated in the actual world. However, the traveller's dying from poisoning is not instantiated in the actual world, so we can conclude that A did not cause the death of the traveller in this case. Mackie, similarly, and for similar reasons, concludes that A did not cause the death of the traveller. He considers a variant of the case where A adds poison to the water and B punctures the canteen. He argues that A does not cause the death of the traveller, since the death was not realised by the causal chain "that starts with poison-in-can", but rather by the causal chain "puncturing-lack-of-water-thirst-death" (46).

We can restate the thought that is common to Wright's and Mackie's arguments in terms of process-connections. While A's poisoning the water is minimally sufficient (in the circumstances and given the laws of nature) for the traveller's death, there is no genuine process-connection between what A did and the death of the traveller. If we add more and more intermediate times between the time at which A poisons the water ( $t_1$ ) and the time at which the traveller dies, we find that there is no such connection. In particular, we find there is no such connection when we consider times after B steals the canteen ( $t_2$ ) but before the traveller dies. Consider any such time  $t_3$ . To connect A's poisoning of the water to the death of the traveller, we would need to add an event D at this time,  $t_3$  – e.g. the traveller's drinking the poisoned water – so that A's poisoning the water belongs to a set of events that is minimally sufficient for D; it would also need to be the case that D in turn belongs to a set of events that is minimally sufficient for the death of the traveller. But there is no such event D in the actual world (remember that minimally sufficiency, just like NESS, states a relation between actual events). Therefore, A's poisoning the water is not process-connected to the death of the traveller.<sup>13</sup>

---

<sup>12</sup> This variant was first introduced by Hart and Honoré (1985[1959]).

<sup>13</sup> There are several ways to explain why A's poisoning the water is not process-connected to the death of the traveller. For one thing, if we think of the traveller's death as it actually occurred, as a death from thirst, not a death from being poisoned, A's poisoning the water is not minimally sufficient for the traveller's death (so understood). It guarantees, not a death from thirst, but at most a death from being poisoned. However, if we do not consider the manner in which the traveller died to be relevant, this explanation is not open to us.

Alternatively, we could explain why A's poisoning the water is not process-connected to the death of the traveller by appealing to time-sensitivity. Most writers discussing this example observe that the traveller's death would have occurred sooner if B had not stolen the canteen, because the traveller would then have been poisoned (e.g. see Stapleton (2008) and Wright (2013)). Since it matters for the timing of the traveller's death whether B steals the canteen or not, it turns out that A's poisoning the water does not belong to a set of events guaranteeing (in the circumstances and given the laws of nature) that the death of the traveller will occur exactly at the time it does. So, A's poisoning the water is not *time-sensitively* minimally sufficient for the traveller's death. (For a general discussion on time sensitivity, see Chapter 12, p. 252ff.)

Since there is no process-connection between A's poisoning the water and the death of the traveller, CAUSATION entails that A's poisoning the water is not a cause of the traveller's death.

Similar remarks can be made about Sartorio's version of the case (in which A replaces the water with sand). In order to connect A's replacing the water with sand to the traveller's death, there needs to be an event D – e.g. the traveller's opening the canteen and finding it full of sand – so that A's replacing the water with sand belongs to a set of events that is minimally sufficient for D, and D in turn belongs to a set of events that is minimally sufficient for the death of the traveller. However, there is no such event D in the actual world. The traveller never finds the canteen full of sand.

At this point, it seems that CAUSATION fails to explain our twin intuitions about THE THIRSTY TRAVELLER. Rather, it seems to entail that neither A nor B caused the death of the traveller. Now we come to the really interesting part. I think Sartorio is on to something when she says that “the man clearly dies as a result of what the two enemies did: after all, he wouldn't have died if they had both refrained from acting in the way they did” (2016: 27). When we say that the traveller would not have not died if they had both refrained from acting in the way they did, we are implicitly (and perhaps unintentionally) setting our focus to the time before any of the agents has acted. At this earlier point in time, there is still a possibility that the traveller will survive; he will survive if neither of his enemies acts in the way they eventually do. At this earlier time, more possibilities are still open. We still have the larger possibility horizon  $H(t_1)$ .

As long as we keep this larger possibility horizon in mind, we will conclude that B's stealing the canteen does cause the death of the traveller. At least, this is what CAUSATION tells us. B's stealing the canteen is process-connected to the traveller's death. It belongs to a set of concurrent events guaranteeing that the traveller's death will occur sometime, that it will occur roughly at the time it did, and that it will occur exactly when it did, and the process-connection remains even if we add more and more intermediate steps (it was not just an apparent process). Moreover, if we remove B's stealing the canteen from this set, the set will no longer guarantee this. So, condition (a) of CAUSATION is satisfied.<sup>14</sup>

Again, given the richer possibility horizon  $H(t_1)$ , B's stealing the canteen makes the traveller's death more secure and his survival less secure (i.e. condition (b) of CAUSATION is satisfied): in the closest possible world where B does not steal the canteen, only one thing would need to change for the death of the traveller not to occur, namely: A's not replacing the water with sand. By contrast, in the actual

---

<sup>14</sup> Mackie (1974) similarly argue that B satisfy the INUS condition for being a cause of the death of the traveller, and Wright (2013) likewise argue that B satisfy the NESS condition for being a cause of the death of the traveller.

world, where B does steal the canteen, an additional thing would need change, namely: B's not stealing the canteen. So, if we consider the larger possibility horizon  $H(t_1)$ , CAUSATION entails that B caused the death of the traveller.<sup>15</sup>

Metaphysically speaking, this is all we can say about what causes what in THE THIRSTY TRAVELLER. It is enough, however, to explain our conflicting intuitions about causation here. If we consider A and B individually, we get the verdict that neither A nor B causes the death of the traveller. A does not cause the death, since his replacing water with sand is not process-connected to the traveller's death, and B does not cause it either, since his stealing the canteen does not increase the security of the death of the traveller given  $H(t_2)$ . However, when we consider A and B together, it is natural to consider it an open possibility that each of them, or both, will act differently. That is, it is natural to apply a possibility horizon like  $H(t_1)$ . And, according to this possibility horizon, it turns out that at least one of them did cause the death of the traveller, namely B.<sup>16</sup>

With this, I conclude that CAUSATION can explain our conflicting intuitions about THE THIRSTY TRAVELLER. In this respect, CAUSATION and Sartorio's appeal to disjunctive facts as causes are on a par. However, CAUSATION can successfully explain our causal intuitions in wide range of other kinds of case, while the appeal to disjunctive facts cannot. For instance, CAUSATION can explain our intuitions in pre-emption cases, but the appeal to disjunctive facts as causes cannot (for the straightforward reason that it does not apply to such cases). This gives us reasons to believe that CAUSATION provides a better explanation of our causal intuitions in THE THIRSTY TRAVELLER than the appeal to disjunctive facts as causes does. After all, an account that can explain our intuitions about this particular case *and* a wide range of other cases is more likely to be correct than one that only explains our intuitions in this particular case.<sup>17</sup>

---

<sup>15</sup> The account of why CAUSATION entails that B causes the death of the traveller given in the main text is imprecise. B does not steal the canteen at  $t_1$  in @. Rather, it is the case that, at  $t_1$  and in @, B will steal the canteen. Still, even if we restate the account in these more precise terms, it follows that CAUSATION entails that B caused the death of the traveller given  $H(t_1)$ .

<sup>16</sup> That said, there could be pragmatic reasons for thinking that the larger possibility horizon is the relevant one in assessing causation, given certain aims we have, and that we therefore have reason to conclude that B caused the death of the traveller. Compare the discussion of which possibility horizon is relevant in the assessment of blameworthiness below.

<sup>17</sup> On top of this, it is doubtful whether facts can be causal relata, and whether facts can be disjunctive. Concerning the latter, Kratzer (2002) argues that facts are particulars. If this is correct, the "Stealing-or-substituting" is not really a fact, but a proposition consisting of two facts. If this is correct, Sartorio could revise her theory to say that propositions might be causes. This, however, seems strange. Still, I will set these questions aside here.

## Who Is to Blame for the Death of the Thirsty Traveller?

Sartorio (2015) argues that A is morally responsible for the death of the traveller because what A did logically entails the disjunctive fact that caused the death of the traveller before what B did does the same. Clearly, this argument presupposes that the causal structure of the case is best explained in terms of disjunctive facts as causes. If my arguments above are cogent, this is not the case, and Sartorio's argument for thinking that A is morally responsible for the death of the traveller fails.

This leaves us with the question of who (if anyone) is morally responsible for the traveller's death. In what remains of this chapter I will use *BLAMEWORTHINESS FOR* to argue that neither A nor B is blameworthy for the death of the traveller, although both of them are blameworthy for trying to bring about his death.

To recap, *BLAMEWORTHINESS FOR* says the following:

*BLAMEWORTHINESS FOR*: you are blameworthy for X rather than X\* just in case there is a Y and Y\*, such that

- (i) X is worse than X\* at least partly since
  - (i1) Y is worse than Y\*, and
  - (i2) there is at least one world in  $H_2$  where X occurs at  $t_2$ , and at least one world in  $H_2$  where X\* occurs at  $t_2$ , and Y is more secure and Y\* is less secure in the closest-to-@-at- $t_2$  world(s) in  $H_2$  where X occurs at  $t_2$  than they are in the closest-to-@-at- $t_2$  world(s) in  $H_2$  where X\* occurs at  $t_2$ .
- (ii) there is a time  $t$ , such that your having a poor quality of will at  $t$  in relation to Y versus Y\* rather the required quality of will is a non-deviant cause of X rather than X\* within the relevant possibility horizon H.<sup>18</sup>

This principle straightforwardly entails that A is not blameworthy for the death of the traveller: since replacing the water with sand is not a cause of the death of the traveller, condition (ii) is not satisfied.<sup>19</sup> If this were a murder trial, we should let A go free now. Or rather, we should charge him with attempted murder. After all, he tried but failed to kill the traveller.

*BLAMEWORTHINESS FOR* also correctly implies that A is blameworthy for replacing the water with sand. (i) A's replacing the water with sand is a bad thing in virtue of the traveller's death being bad. The death of the traveller is more secure given that

---

<sup>18</sup> This is one of the elaborated versions presented towards the end of Chapter 12.

<sup>19</sup> To be precise, condition (ii) of *BLAMEWORTHINESS FOR* is not satisfied since A's having a poor quality of will towards the traveller is not a cause of the traveller's death. The reason for this is the same as the reason why A's replacing the water with sand does not cause traveller's death.



A replaces the water with sand. (ii) There is a time  $t$ , such that A's poor quality of will in relation to the traveller is a non-deviant cause of A's replacing the water with sand, within the relevant possibility horizon. At least, this is what CAUSATION entails. (a) His having a poor quality of will is process-connected to his replacing the water with sand, and (b) his replacing the water with sand is more secure in @ than in the closest-to-@-at- $t$  world where he has the required quality of will. Moreover, his quality of will seems to be a non-deviant cause of his actions at this time. There is, for instance, no evil neuroscientist intervening in his thought processes. So, BLAMEWORTHINESS FOR entails that A is blameworthy for attempting to murder the traveller, but not for murdering him.

Turning to B, given the smaller possibility horizon  $H(t_2)$  (shown on p. 312) representing the fact that there is no possibility that the traveller will survive, BLAMEWORTHINESS FOR entails that B, like A, is not to blame for the death of the traveller. With this possibility horizon, B does not cause the death of the traveller, and condition (ii) of BLAMEWORTHINESS FOR is therefore not satisfied.

However, we have several pragmatic reasons for thinking that the larger possibility horizon is the relevant one for assessing blameworthiness in this case. First, when we are assessing blameworthiness, we have reasons to consider the quality of will of all the agents involved in the situation. This is what RELEVANT POSSIBILITIES FOR BLAME tells us. If we do not consider it a possibility that A and B could have had the required quality of will in relation to the traveller, we will assume that their poor qualities of will did not cause the outcome, and by extension we will affirm their innocence. And, perhaps needless to say, an affirmation of the innocence of A and B seems the wrong place to start when we are assessing whether or not they are blameworthy.

Second, even if we were not in the business of assessing blameworthiness, the reduced possibility horizon  $H(t_2)$  is unsatisfactory. In this possibility horizon, the traveller's death protrudes as an uncaused event. If it is important to us to decide what brought about the death, then whatever our interest is (perhaps it is to identify the cause of the traveller's death in order to prevent future travellers dying under similar circumstances) we will surely have good reason to look for, and consider carefully, a range of candidate causes. Hence, we have good reason to widen and adjust the possibility horizon we are currently operating with (in line with the principle WIDENING AND ADJUSTING suggested in "You Just Didn't Care Enough").<sup>20</sup>

Given the relevant larger possibility horizon  $H(t_2)$ , BLAMEWORTHINESS FOR entails that B might be blameworthy for the death of the traveller. The death of the traveller just is worse than his survival. So, condition (i) is satisfied. And given this

---

<sup>20</sup> The paper referred to here, presented in Chapter 11, discusses why the richer possibility horizon is typically the relevant one (see p. 237ff).

possibility horizon, B's having a substandard quality of will in relation to the traveller's dying rather than surviving caused the traveller to die rather than survive.

Still, I think the correct verdict is that B is not blameworthy for the death of the traveller. Although B's poor quality of will towards the traveller caused the traveller's death in this possibility horizon, it did so in a deviant way. Bernstein (2019) suggests this, albeit in relation to a modified version of the case like Mackie's and Wright's where A adds poison to the canteen.<sup>21</sup> I think the same claim can be made about the version I have been considering here. If this idea holds up to scrutiny, THE THIRSTY TRAVELLER is interesting in a perhaps unexpected way.

Remember, NON-DEVIANT CAUSE says that in cases where you have an intention to X (perform an action, omit doing something or bring about an outcome) your quality of will is a non-deviant cause of X if and only if (i) X is brought about roughly as you had planned, and (ii) causation does not go astray before you form your intention.<sup>22</sup> While the death of the traveller is brought about roughly as B planned, it seems that causation did go astray, in this case, before B formed his intention. One indication of this is that B would probably not have formed an intention to steal the canteen if he had known that it was filled with sand. Or, if he already had already formed the intention to steal the canteen when he found out that it was filled with sand, he would probably have changed his mind and not stolen it. If this is correct, THE THIRSTY TRAVELLER is what Mele (1987) calls a case of *tertiary waywardness*: the causal chain goes astray even though the outcome is brought about roughly as planned.

Still, B does not emerge as a completely blameless agent. It seems we can blame him for attempting murder at least. This is also what BLAMEWORTHINESS FOR entails. It entails that B is blameworthy for stealing the canteen (and thereby attempting to kill the traveller) because, roughly, (i) stealing the canteen is worse than not doing so in virtue of the fact the stealing makes the death of the traveller more secure than it otherwise would have been, and (ii) B's poor quality of will towards the traveller is a non-deviant cause of the death of the traveller. We see this if we consider CAUSATION. (a) B's poor quality of will is process-connected to the death of the traveller, and (b) it makes the death of the traveller more secure. As a final point, I think we can safely say that B's poor quality of will was a non-deviant cause of his stealing the canteen. No evil neuroscientist is lurking in the background

---

<sup>21</sup> Bernstein (2019) also suggests that A causes the death of the traveller in a deviant way. I do not agree, since I do not think that A causes the death of the traveller at all. Still, if it turns out that I am wrong in thinking this, Bernstein is correct. The death of the traveller is not brought about in the way A envisaged. Thus, he would not find that the canteen is filled with sand, but instead find no canteen at all, when he becomes thirsty in the desert.

<sup>22</sup> See discussion in Chapter 12, p. 246ff.

here. Thus, we can conclude that BLAMEWORTHINESS FOR entails that B is blameworthy for stealing the canteen.<sup>23</sup>

## Conclusion

To sum up, we have not yet seen a convincing counterexample to the idea that you are only morally responsible for an outcome if you cause it. The arguments Sartorio presents in challenging this idea fail. Sometimes, she wrongly assumes that replacing an agent with a non-agential phenomenon is always causally irrelevant, as when she compares TWO BUTTONS with ONE BUTTON, ONE MECHANISM. At other times, she overlooks causally relevant details, such as the fact that B's bad intentions make a difference to whether the traveller will die in THE THIRSTY TRAVELLER WITH A FRIEND.

At this point, we might be inclined to conclude that MORAL ENTAILS CAUSAL is correct. However, ultimately, whether an agent can be morally responsible for an outcome without causing it depends on which accounts of moral responsibility and causation are correct. It is easy to show that MORAL ENTAILS CAUSAL is sound if BLAMEWORTHINESS FOR is correct. BLAMEWORTHINESS FOR states, in essence, that you are blameworthy for some outcome X only if your  $\varphi$ -ing is a *cause* of X (see condition (ii)).

Although BLAMEWORTHINESS FOR requires us to think about causation along the lines of CAUSATION, it does not matter, for the acceptability of MORAL ENTAILS CAUSAL, which account of causation is correct. As long as we apply the same account of causation when deciding whether someone is blameworthy for an outcome as we do when deciding whether that person caused it, moral will entail causal.

As a final point, it is worth repeating that we have found that both of the enemies of the thirsty traveller are guilty of attempted murder, but that they are so for different reasons. This is because the first enemy did not cause the traveller's death, while the second did albeit via a deviant causal process.

---

<sup>23</sup> It might be argued that both A and B are blameworthy in the sense that they are bad persons. As Talbert (2015) suggests: "assessments of moral responsibility are ultimately attuned (or at least ought to be) to our perception of how a person is as a moral agent and how she is oriented toward other moral agents" (181). I will not go into this here, but I think there is an interesting sense in which this is correct. However, as I argued in Chapter 10, the question of whether A and B are bad persons is orthogonal to the question of whether they are blameworthy for (here) the death of the traveller (or for trying to bring it about).

# Reasons, Blame, and Collective Harms

I started out asking under what conditions you have outcome-related reasons to act in the relevant way  $\varphi$  in collective harm cases, and ended up proposing, with Caroline Touborg, both a new account of outcome-related reasons (called REASON), and a new account of when you are blameworthy for X, where X is an action, omission or outcome (called BLAMEWORTHINESS FOR). According to these accounts, roughly, you have an outcome-related reason to  $\varphi$  if and only if  $\varphi$ -ing makes some good outcome more secure within the relevant possibility horizon, and you are blameworthy for X if and only if X is bad and your poor quality of will towards X was a non-deviant cause of it.

To give an example, REASON says, essentially, that you have a reason not to cross a beautiful lawn if and only if it is possible that the lawn will be ruined (if enough people cross it), it is possible that it will not be ruined (if enough people instead go around it), and your crossing it takes us closer to its being ruined. And, BLAMEWORTHINESS FOR says, roughly, that you are blameworthy for ruining the lawn if and only if it is bad that the lawn is ruined, and your disregard for the beauty of the lawn was a non-deviant cause of its being ruined. To give some additional examples, these principles entail that you have a reason to vote for the right candidate in an election, and that you might be blameworthy if you do not (given that there is a possibility that the candidate will win, and a possibility that he or she will not), that you have a reason not to buy factory-farmed chickens, and that you might be blameworthy if you do (given that there is a possibility that chickens will suffer, and a possibility that they will not), and so on.

Both REASON and BLAMEWORTHINESS FOR builds on a particular account of causation, namely Touborg's (2018) account according to which an event C is a cause of another event E if and only if C is process-connected to E, and C makes E more secure within the relevant possibility horizon. I have called this account CAUSATION. The requirement that a cause must be process-connected to its outcome captures the idea that a cause must be involved in the causal history of its outcome. In turn, the requirement that a cause must make its effect more secure captures the idea that a cause contributes to its effect. REASON makes use only of the second requirement in stating, in essence, that you have reasons to contribute to good outcomes. Conversely, BLAMEWORTHINESS FOR makes use of both requirements. It states that you are blameworthy for X only if you are involved in the causal history of X, and contribute to X.

The major virtue of these accounts is that they give intuitively correct verdicts about what reasons we have, and about whether somebody is blameworthy for some action, omission or outcome, in many different kinds of cases. They give intuitively correct verdicts in collective harm cases with a threshold, collective harm cases without a threshold, pre-emption cases, switching cases, omission cases, Frankfurt-style cases, cases where we disregard irrelevant possibilities, the difficult case of the thirsty traveller, and more. In addition, they can explain the twin intuitions about reasons and blameworthiness we might have in some of these cases. Take a threshold case like voting. They can explain why it seems that you have a reason to vote for the best candidate in an election if there is a possibility that the candidate will win and if your voting can contribute to the win, while it also seems that you lack a reason to vote for the candidate because your vote will not make a difference for whether this candidate wins or not given that others vote as they do. These different verdicts stem from different ways of thinking about the situation; they stem from different possibility horizons.

In many cases where we get different verdicts about reasons or blameworthiness when considering different possibility horizons, there are pragmatic reasons for thinking that the larger possibility horizon is the more correct one. For example, consider a case like the following: I did not vote for the appropriate party in a parliamentary election; in fact, I did not bother to vote at all. Further, as things turned out, many of us who would have voted for the appropriate candidate did not bother voting, with the result that the worst party won the election with some margin. If this was the case, you might accuse me for not bothering to vote, saying that this contributed to the horrible outcome. However, I might defend myself, saying “given that many others did not bother to vote, my not voting did not make a difference”. Considered on its own, you might find this defence acceptable. However, if the other idlers also defend themselves in this way, you will probably find our defences questionable. For one thing, these defences do not fit well together. Each of us is in effect saying that he could have acted otherwise, but asking you to treat what the other idlers did as fixed. So, if you accept our defences, you must both treat it an open possibility that each of us could have acted otherwise, *and* treat it as fixed that each of us acted as he did. This is contradictory. For another, each of us is essentially saying that whether he voted or not did not matter for which party that won the election. Still, taken together, it seems that it *did* matter for who won the election that each of us did not bother to vote. So, if you accept our defences, you cannot easily explain why it seems that the worst candidate won the election because we did not bother to vote. This provides you with another reason to reconsider our defences.

Collective harm cases without a threshold (“non-threshold cases”) often play a pivotal role in discussions about what reasons we have in collective harms cases. Much ink has been spilled to show that, upon closer scrutiny, there are no genuine non-threshold cases. I think that some of these arguments succeed. In alleged non-

threshold cases, each act of  $\varphi$ -ing does makes a morally relevant difference. These differences might be so small that they are imperceptible, but they are still morally relevant. For instance, in Parfit's (1984) case DROPS OF WATER (see p. 33), each drop of water makes a morally relevant difference to the people's suffering although no drop of water makes a perceptible difference in thirst for anyone.

Yet, even if it turns out that these arguments fail and that genuine non-threshold cases are possible, REASON and BLAMEWORTHINESS FOR give the right verdicts about what reasons we have, and about who is blameworthy for what. Drawing on contemporary accounts of vagueness, Touborg and I argue that acting in the relevant way ( $\varphi$ ) makes the outcome more secure also in non-threshold cases. The idea is this. In order for non-threshold cases to be possible, the outcome must be vague. (If it is not vague, it has sharp boundaries, which in turn means that one act of  $\varphi$ -ing could make a difference for whether the outcome occurs or not, and so it turns out that the case under consideration is not a genuine non-threshold case after all.) Contemporary accounts of vagueness indicate, however, that outcomes like the alleviation of the people's suffering is not best described as one big event with imprecise conditions of occurrence. Instead, there are many candidates for being the right description of the alleviation of the people's suffering, each with precise conditions of occurrence, but nobody has been fool enough to try to pinpoint which candidate is the correct one. Given that the alleviation of the people's suffering is best described as an event with precise conditions of occurrence, donating one pint of water could make a difference for whether this precisely described event will occur. Then, if you treat it as an open possibility that the people's suffering will be alleviated, and a possibility that their suffering will not be alleviated, REASON entails that you have reason to donate your pint because doing so makes it more secure that suffering will be alleviated rather than continuing unchanged. This is true even if donating a pint makes no perceptible difference in thirst for anyone, and even if only perceptible differences matter morally. Using a similar kind of reasoning, BLAMEWORTHINESS FOR entails that you are blameworthy for the people's continued suffering if you keep your pint for yourself and enough others do the same.

Throughout most of the thesis, I have treated climate change as a non-threshold case. I have assumed that, for any small-scale emissions generating event like going for a ride in a fossil fuel powered car, it is true that climate change will occur just the same whether or not this emissions generating event occurs or not. Still, it is far from sure that climate change is a non-threshold case. Broome (2019) argues for instance that empirical evidence indicates that, because of the atmosphere's instability, each drive with a fossil fuel driven car makes a difference for future climate change and its related harms. If this is correct, climate change is best described as a well-known case of choice under uncertainty. It is a case where each act makes a difference for the outcome, and where we cannot know whether it will make a difference for better or worse, but where we have good grounds for thinking

that, on average, each act makes a difference for the worse. Climate change has also been described as an overdetermination case (Cripps 2013), a pre-emption case (Lawford-Smith 2016; Eriksson 2019), and a collective impact case with a threshold (Kagan 2011). In the end, I think that REASON and BLAMEWORTHINESS FOR can explain our intuitions about what reasons we have not to contribute to climate change, and about who is blameworthy for contributing to climate change, because they can explain our intuitions in all these kinds of cases.

The initial motivation for developing REASON and BLAMEWORTHINESS FOR was to explain where the inefficacy argument goes wrong. One version of this argument says that you have no outcome related reason to  $\varphi$  in collective harm cases since the outcome will occur whether you  $\varphi$  or not. More elaboratively, this version of the inefficacy argument says the following:

- (i) If outcome O will occur whether you  $\varphi$  or not, you have no O-related reason to  $\varphi$ .
  - (ii) Outcome O will occur whether you  $\varphi$  or not.
- ∴ You have no O-related reason to  $\varphi$ .

This is the version of the argument I have focused on throughout the thesis. We are now equipped to pinpoint where this argument goes wrong. REASON entails that premise (i) is incorrect. You might have an O-related reason to  $\varphi$  even if outcome O will occur whether you  $\varphi$  or not. You have such a reason if O is a good outcome and  $\varphi$ -ing makes O more secure within the relevant possibility horizon. In more general terms: you have such a reason if  $\varphi$ -ing contributes to some good outcome O. So, to the extent we have reason to believe in REASON, premise (i) is incorrect.

You could however understand the inefficacy argument in a different way. Another version of this argument says that you have no outcome related reason to  $\varphi$  in collective harm cases since  $\varphi$ -ing makes no difference at all to the outcome. This version of the argument can be specified as follows:

- (i) If  $\varphi$ -ing makes no difference to outcome O, you have no O-related reason to  $\varphi$ .
  - (ii)  $\varphi$ -ing makes no difference to outcome O.
- ∴ You have no O-related reason to  $\varphi$ .

If this is how you understand the inefficacy argument, premise (ii) will often be incorrect. At least, this is what REASON tells us. Premise (ii) will be incorrect in collective harm cases when it is up in the air whether O will occur. That is, (ii) will be incorrect if there is a relevant possibility that O will occur and a relevant possibility that O will not occur at the time at which you could  $\varphi$ . In that case,  $\varphi$ -ing

makes a kind of difference to O, it makes it more secure. It brings O closer to happening (or further from not happening). Premise (ii) might of course hold in other kinds of cases. In switching cases such as THE ENGINEER, where there is no possibility that the outcome will fail to occur,  $\varphi$ -ing does not make any difference to the occurrence of the outcome – not even to the security of the outcome – which entails that you have no outcome-related reason to  $\varphi$ .

At the outset, I did not give any precise description of outcome-related reasons. I merely tried to give an idea of what they are by way of giving examples. I said that they are climate-change-related reasons not to joy-guzzle, future-suffering-of-chickens-related reason not to buy factory-farmed chicken, and so on. At that point, I wanted to keep it open how we should understand these reasons. It might have turned out that they are grounded in considerations of fairness, Kantian duties, collectivization duties, or something else completely. Now, however, I think we are in a position to say that outcome-related reasons are best described as teleological reasons, also called consequentialist or instrumental reasons. These are reasons that speaks in favour of some action in virtue of the fact that this action contributes to some outcome. If REASON is correct, we can be even more precise. We can say that outcome-related reasons speak in favour of some action in virtue of the fact that this action makes some worthy outcome more secure.

Conversely, at the outset of Part Two, I described quite elaboratively what it is to be blameworthy for something. First, following Strawson (2008/1962) and others, I took it that being blameworthy means that you are the appropriate target of reactive attitudes like resentment, indignation, guilt, and so on. Second, I suggested that there is a distinction to be made between *blameworthiness* and *blameworthiness for*, where assessments of *blameworthiness* are evaluations of the character or quality of will of a person (perhaps in combination with some further condition), whereas assessments of who is *blameworthy for* something are assessments of whose fault this something is, if anyone's.

In addition, I argued *pace* Zimmerman (2002) and others that the question whether you are blameworthy for an outcome is not otiose. Whether you are blameworthy for an outcome might influence the degree to which others are warranted in resenting you, be indignant about what you did, and so on. If this is correct, there is such a thing as resultant moral luck. This idea is controversial. Still, even if it turns out that I am mistaken on this point, we cannot conclude that questions about whether you are blameworthy for an outcome are otiose. Discussions about who's fault something is are important and abundant in ordinary life as well as in a scholarly contexts, and I expect that they will be so even if settling these matters are irrelevant for deciding how blameworthy people are.



## Questions for Future Research

Even though I hope to have given comprehensive enough arguments for the conclusions of this thesis, some questions are left for future research. First, there are plenty of accounts of reasons, blameworthiness and causation other than those I have been able to evaluate in this thesis. To say something about them, and in particular how they relate to REASON, BLAMEWORTHINESS FOR and CAUSATION, will have to wait for another day.

Second, BLAMEWORTHINESS FOR is a compatibilist account of moral responsibility. It entails that you might be blameworthy for what you do in a deterministic world. There are several well-known arguments against the compatibilist thesis, such as the manipulation argument and the consequence argument. These arguments have already been extensively discussed elsewhere, but it would have been interesting to see whether we could use the resources of BLAMEWORTHINESS FOR to disarm them.

Third, it is commonly assumed that there are two conditions on blameworthiness: a control condition and an epistemic condition. I have not said much about the role knowledge plays in matters of blameworthiness. I think it does play a role, and that BLAMEWORTHINESS FOR can explain this role. To say more about this issue will also have to be a topic for another occasion.

Fourth, there is a discussion on what it means to harm someone that closely parallels the discussion on what it is for an event to cause another. Several of the most prominent accounts of what it means to harm someone builds on the idea that you harm someone if and only if this person would have been better off if you had acted differently. That is, these accounts essentially incorporate something like a simple counterfactual account of causation (SIMPLE). For this reason, they also face counterexamples in the form of pre-emption and overdetermination cases. It seems to me that we could avoid these counterexamples if we draw on the resources of CAUSATION. That is, they could be avoided if we build our account of harm on a weaker form of counterfactual dependence, namely security dependence. We could for instance say that you harm someone if and only if you cause this person harm, where causation is understood along the lines of CAUSATION. Even better, since harm often is taken to be a contrastive notion, we should include causal contrasts in our account, as follows: you harm someone if and only if you cause this person to be worse off than she otherwise would have been. Still, this idea needs to be refined in several ways to be successful. For instance, harm is often taken to be not only a causal relation between events, but also a constitute relation. Certain kinds of events in our nervous system does, for example, constitute harm, or ground harm, even though they are not harms themselves. This idea is not easily explained using an account of causation.

# References

- Algander, Per (2013) *Harm, benefit, and non-identity* (Doctoral thesis). Uppsala University.
- Andersson, Henrik (2017) *How it all relates: Exploring the space of value comparisons* (Doctoral thesis). Lund University.
- Annas, Julia (2011) *Intelligent virtue*. Oxford: Oxford University Press.
- Armstrong, D. M. (1968) *A materialist theory of the mind*. London: Routledge & Kegan Paul.
- Arneson, Richard J. (1982) "The principle of fairness and free-rider problems". *Ethics*, 92(4): 616-33.
- Arntzenius, Frank, & David McCarthy (1997) "Self torture and group beneficence". *Erkenntnis*, 47(1): 129-44.
- Aronson, Jerrold L. (1971) "On the grammar of 'cause'". *Synthese*, 22(3/4): 414-30.
- Baatz, Christian (2014) "Climate change and individual duties to reduce GHG emissions". *Ethics, Policy & Environment*, 17(1): 1-19.
- Banks, Melany (2010) *Individual responsibility for collective harms* (Doctoral thesis). Wilfrid Laurier University.
- Barnes, Elizabeth (2010) "Ontic vagueness: A guide for the perplexed". *Nous*, 44(4): 601-27.
- Barnett, Zach (2018) "No free lunch: The significance of tiny contributions". *Analysis*, 78(1): 3-13.
- Beebe, Helen (2004) "Causing and nothingness" in J. Collins, N. Hall, & L. A. Paul (Eds.) *Causation and counterfactuals* (291-308). Cambridge, Mass.: MIT Press.
- Bennett, Daniel (1965) "Action, reason, and purpose". *The Journal of Philosophy*, 64(4): 85-96.
- Bernstein, Sara (2019) "Moral luck and deviant causation". *Midwest Studies In Philosophy*, 43(1): 151-61.
- Bishop, John (1989) *Natural agency: An essay on the causal theory of action*. Cambridge: Cambridge University Press.
- Björnsson, Gunnar (2011) "Joint responsibility without individual control: Applying the explanation hypothesis" in J. van den Hoven, I. van de Poel, & N. Vincent (Eds.) *Moral responsibility: Beyond free will and determinism*. Dordrecht: Springer.
- Björnsson, Gunnar (2014) "Essentially shared obligations". *Midwest Studies In Philosophy*, 38(1): 103-20.

- Björnsson, Gunnar (2017a) "Explaining (away) the epistemic condition on moral responsibility" in P. Robichaud & J. W. Wieland (Eds.) *Responsibility: The epistemic condition* (146-62). Oxford: Oxford University Press.
- Björnsson, Gunnar (2017b) "Explaining away epistemic skepticism about culpability" in S. David (Ed.), *Oxford studies in agency and responsibility* (Vol. 4: 141–64). Oxford: Oxford University Press.
- Björnsson, Gunnar (2021) "On individual and shared obligations: In defense of the activist's perspective" in M. Budolfson, T. McPherson, & D. Plunkett (Eds.) *Philosophy and climate change*. Oxford: Oxford University Press.
- Braham, Matthew, & Martin Van Hees (2012) "An anatomy of moral responsibility". *Mind*, 121(483): 601-34.
- Brand, Myles (1984) *Intending and acting: Toward a naturalized action theory*. Cambridge, Mass.: MIT Press.
- Bratman, Michael (1993) "Shared intention". *Ethics*, 104(1): 97-113.
- Bratman, Michael (2014) *Shared agency: A planning theory of acting together*. New York: Oxford University Press.
- Brennan, Jason (2009) "Polluting the polls: When citizens should not vote". *Australasian Journal of Philosophy*, 87(4): 535-49.
- Brink, David O, & Dana Nelkin (2013) "Fairness and the architecture of responsibility" in D. Shoemaker (Ed.), *Oxford studies in agency and responsibility* (Vol. 1: 284-313). Oxford: Oxford University Press.
- Broome, John (2004) *Weighing lives*. Oxford: Oxford University Press.
- Broome, John (2012) *Climate matters: Ethics in a warming world*. New York: W.W. Norton & Company.
- Broome, John (2019) "Against denialism". *The Monist*, 102(1): 110-29.
- Budolfson, Mark Bryant (2019) "The inefficacy objection to consequentialism and the problem with the expected consequences response". *Philosophical Studies*, 176(7): 1711-24.
- Bunzl, Martin (1979) "Causal overdetermination". *Journal of Philosophy*, 76(3): 134-50.
- Cargile, James (1969) "The sorites paradox". *The British Journal for the Philosophy of Science*, 20(3): 193-202.
- Carlson, Erik (1996) "Cyclical preferences and rational choice". *Theoria*, 62(1/2): 144-60.
- Carlson, Erik, Magnus Jedenheim-Edling, & Jens Johansson (2021) "The significance of tiny contributions: Barnett and beyond". *Utilitas*: 1-9.
- Chang, Ruth (2002) "The possibility of parity". *Ethics*, 112(4): 659-88.
- Chisholm, Roderick M. (1964a) "The descriptive element in the concept of action". *Journal of Philosophy*, 61(20): 613-25.
- Chisholm, Roderick M. (1964b) "Human freedom and the self". *The Lindley Lecture*. Lawrence: University of Kansas
- Clarke, Randolph (2011) "Omissions, responsibility, and symmetry". *Philosophy and Phenomenological Research*, 82(3): 594-624.

- Clarke, Randolph (2014) *Omissions: Agency, metaphysics, and responsibility*. New York: Oxford University Press.
- Collins, Stephanie (2019) *Group duties: Their existence and their implications for individuals*. Oxford: Oxford University Press.
- Cripps, Elizabeth (2013) *Climate change and the moral agent: Individual duties in an interdependent world*. Oxford: Oxford University Press.
- Cullity, Garrett (2000) "Pooled beneficence" in M. J. Almeida (Ed.), *Imperceptible harms and benefits* (1-23). Dordrecht: Kluwer Academic Publishers.
- Cullity, Garrett (2019) "Climate harms". *The Monist*, 102(1): 22-41.
- Davidson, Donald (1963) "Actions, reasons, and causes". *The Journal of Philosophy*, 60(23): 685-700.
- Dowe, Phil (2000) *Physical causation*. Cambridge: Cambridge University Press.
- Driver, Julia (2001) *Uneasy virtue*. Cambridge: Cambridge University Press.
- Dummett, Michael (1975) "Wang's paradox". *Synthese*, 30(3/4): 201-32.
- Enoch, David, & Andrei Marmor (2007) "The case against moral luck". *Law and Philosophy*, 26(4): 405-36.
- Eriksson, Anton (2019) *Omitting to emit: Moral duties to reduce emissions in global supply chains* (Doctoral thesis). University of Sheffield.
- Fanciullo, James (2020) "What is the point of helping?". *Philosophical Studies*, 177(6): 1487-500.
- Feinberg, Joel (1970) "Sua culpa" in *Doing & deserving: Essays in the theory of responsibility*. Princeton: Princeton University Press.
- Feinberg, Joel (1984) *Harm to others*. New York: Oxford University Press.
- Feit, Neil (2015) "Plural harm". *Philosophy and Phenomenological Research*, 90(2): 361-88.
- Feldman, Fred (1980) "The principle of moral harmony". *Journal of Philosophy*, 77(3): 166-79.
- Fenton-Glynn, Luke (2017) "A proposed probabilistic extension of the halpern and pearl definition of 'actual cause'". *The British Journal for the Philosophy of Science*, 68(4): 1061-124.
- Fine, Kit (1975) "Vagueness, truth and logic". *Synthese*, 30(3): 265-300.
- Fischer, John Martin (2006) "The free will revolution (continued)". *The Journal of Ethics*, 10(3): 315-45.
- Fischer, John Martin, & Mark Ravizza (1998) *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.
- Foot, Philippa (1967) "The problem of abortion and the doctrine of double effect". *Oxford Review*, 5: 5-15.
- Frankfurt, Harry G. (1969) "Alternate possibilities and moral responsibility". *Journal of Philosophy*, 66(23): 829.
- Frankfurt, Harry G. (1971) "Freedom of the will and the concept of a person". *Journal of Philosophy*, 68(1): 5-20.

- French, Peter (1984) *Collective and corporate responsibility*. New York: Columbia University Press.
- Fumerton, Richard, & Ken Kress (2001) "Causation and the law: Preemption, lawful sufficiency, and causal sufficiency". *Law and Contemporary Problems*, 64(4):83-105.
- Gavison, Ruth, Avishai Margalit, & Edna Ullmann-Margalit (1980) "Causal overdetermination: Resolution of a puzzle". *Philosophical Studies*, 37(4): 381-90.
- Gilbert, Margaret (1989) *On social facts*. New York: Routledge.
- Gilbert, Margaret (2014) *Joint commitment: How we make the social world*. New York: Oxford University Press.
- Ginet, Carl (2006) "Working with Fischer and Ravizza's account of moral responsibility". *The Journal of Ethics*, 10(3): 229-53.
- Glover, Jonathan, & M. J. Scott-Taggart (1975) "It makes no difference whether or not I do it". *Proceedings of the Aristotelian Society*, 49: 171-209.
- Glynn, Luke (2011) "A probabilistic analysis of causation". *The British Journal for the Philosophy of Science*, 62(2): 343-92.
- Goodman, Nelson (1951) *The structure of appearance*. Cambridge, Mass.: Harvard University Press.
- Graff, Delia (2001) "Phenomenal continua and the sorites". *Mind*, 110(440): 905-35.
- Gruzalski, Bart (1986) "Parfit's impact on utilitarianism". *Ethics*, 96(4): 760-83.
- Gunnemyr, Mattias (2019) "Causing global warming". *Ethical Theory and Moral Practice*, 22(2): 399-424.
- Gunnemyr, Mattias (2020) "Why the social connection model fails: Participation is neither necessary nor sufficient for political responsibility". *Hypatia*, 35(4): 567-86.
- Hall, Ned (2000) "Causation and the price of transitivity". *Journal of Philosophy*, 97(4): 198-222.
- Hall, Ned (2004) "Two concepts of causation" in J. Collins, N. Hall, & L. A. Paul (Eds.) *Causation and counterfactuals* (225-76). Cambridge, Mass.: MIT Press.
- Hall, Ned (2007) "Structural equations and causation". *Philosophical Studies*, 132(1): 109-36.
- Halldén, Sören (1949) *The logic of nonsense* (Vol. 2). Uppsala: Uppsala Universitets Arsskrift.
- Hanna, Nathan (2014) "Moral luck defended". *Nous*, 48(4): 683-98.
- Hart, H. L. A. (1955) "Are there any natural rights?". *Philosophical Review*, 64(2): 175-91.
- Hart, H. L. A., & Tony Honoré (1985) *Causation in the law* (2nd ed.). Oxford: Clarendon Press.
- Hartman, Robert J. (2017) *In defense of moral luck: Why luck often affects praiseworthiness and blameworthiness*. New York: Routledge.
- Hedden, Brian (2020) "Consequentialism and collective action". *Ethics*, 130(4): 530-54.
- Held, Virginia (1970) "Can a random collection of individuals be morally responsible?". *Journal of Philosophy*, 67(14): 471-81.
- Hieronymi, Pamela (2008) "Responsibility for believing". *Synthese*, 161(3): 357-73.

- Hill, Thomas E. (1983) "Ideals of human excellence and preserving natural environments". *Environmental Ethics*, 5(3): 211-24.
- Hiller, Avram (2011) "Climate change and individual responsibility". *The Monist*, 94(3): 349-68.
- Hindriks, Frank (2019) "The duty to join forces: When individuals lack control". *The Monist*, 102(2): 204-20.
- Hitchcock, Christopher (2001) "The intransitivity of causation revealed in equations and graphs". *Journal of Philosophy*, 98(6): 273-99.
- Hobbes, Thomas (1997/1651) *Leviathan* (R. E. Flathman & D. Johnston Eds.). New York: Norton.
- Holtug, Nils (1998) "Egalitarianism and the levelling down objection". *Analysis*, 58(2): 166-74.
- Hormio, Såde (2017) *Marginal participation, complicity, and agnotology: What climate change can teach us about individual and collective responsibility* (Doctoral thesis). University of Helsinki.
- Hume, David (1999/1748) *An enquiry concerning human understanding*. Oxford: Oxford University Press.
- Hume, David (2007/1738-40) *A treatise of human nature: A critical edition*. Oxford: Clarendon.
- Hursthouse, Rosalind (1991) "Virtue theory and abortion". *Philosophy and Public Affairs*, 20(3): 223-46.
- Hursthouse, Rosalind (1999) *On virtue ethics*. Oxford: Oxford University Press.
- Isaacs, Tracy (2011) *Moral responsibility in collective contexts*. New York: Oxford University Press.
- Jackson, Frank (1997) "Which effects" in J. Dancy (Ed.), *Reading Parfit* (42-53). Oxford: Blackwell.
- Jackson, Frank, & R. J. Pinkerton (1973) "On an argument against sensory items". *Mind*, 82(326): 269-72.
- Jamieson, Dale (2007) "When utilitarians should be virtue theorists". *Utilitas*, 19(2): 160-83.
- Jeppsson, Sofia (2016) "Reasons, determinism and the ability to do otherwise". *Ethical Theory and Moral Practice*, 19(5): 1225-40.
- Johansson, Jens, & Olle Risberg (2019) "The preemption problem". *Philosophical Studies*, 176(2): 351-65.
- Johnson, Baylor (2003) "Ethical obligations in a tragedy of the commons". *Environmental Values*, 12(3): 271-87.
- Johnson, Baylor (2011) "The possibility of a joint communique: My response to Hourdequin". *Environmental Values*, 20: 147-56.
- Kagan, Shelly (2011) "Do I make a difference?". *Philosophy & Public Affairs*, 39(2): 105-41.
- Kant, Immanuel (2002/1785) *Groundwork for the metaphysics of morals* (A. W. Wood, Trans.). New Haven: Yale University Press.

- Keefe, Rosanna (2000) *Theories of vagueness*. Cambridge: Cambridge University Press.
- Kingston, Ewan, & Walter Sinnott-Armstrong (2018) "What's wrong with joyguzzling?". *Ethical Theory and Moral Practice*, 21(1): 169-86.
- Korsgaard, Christine M. (1996) *Creating the kingdom of ends*. Cambridge: Cambridge University Press.
- Kratzer, Angelika (2002) "Facts: Particulars or information units?". *Linguistics and Philosophy*, 25(5-6): 655-70.
- Kutz, Christopher (2000) *Complicity: Ethics and law for a collective age*. New York: Cambridge University Press.
- Kvart, Igal (2002) "Probabilistic cause and the thirsty traveler". *Journal of Philosophical Logic*, 31(2): 139-79.
- Lane, Melissa (2018) "Uncertainty, action and politics: The problem of negligibility" in K. Forrester & S. Smith (Eds.) *Nature, action and the future: Political thought and the environment* (157-79). Cambridge: Cambridge University Press.
- Lawford-Smith, Holly (2015) "Unethical consumption & obligations to signal". *Ethics and International Affairs*, 29(3): 315-30.
- Lawford-Smith, Holly (2016) "Difference-making and individuals' climate-related obligations" in C. Heyward & D. Roser (Eds.) *Climate justice in a non-ideal world* (64-82). Oxford: Oxford University Press.
- Lawson, Brian (2013) "Individual complicity in collective wrongdoing". *Ethical Theory & Moral Practice*, 16(2): 227-43.
- Levy, Neil (2005) "The good, the bad, and the blameworthy". *Journal of Ethics and Social Philosophy*, 1(2): 1-16.
- Lewis, David (1973a) "Causation". *The Journal of Philosophy*, 70(17): 556-67.
- Lewis, David (1973b) *Counterfactuals*. Oxford: Blackwell.
- Lewis, David (1986a) "Causation (with postscript)" in *Philosophical papers vol 2* (159-213). New York: Oxford University Press.
- Lewis, David (1986b) "Events" in *Philosophical papers vol 2* (241-69). New York: Oxford University Press.
- Lewis, David (1986c) *On the plurality of worlds*. Oxford: Blackwell.
- Lewis, David (2000) "Causation as influence". *Journal of Philosophy*, 97(4): 182-97.
- Lewis, David (2004) "Causation as influence (extended)" in J. Collins, N. Hall, & L. A. Paul (Eds.) *Causation and counterfactuals* (75-106). Cambridge, Mass.: MIT Press.
- List, Christian, & Philip Pettit (2011) *Group agency: The possibility, design, and status of corporate agents*. Oxford: Oxford University Press.
- Ludwig, Kirk (2016) *From individual to plural agency: Collective action* (Vol. 1). Oxford: Oxford University Press.
- Ludwig, Kirk (2017) "Methodological individualism, the we-mode, and team reasoning" in G. Preyer & G. Peter (Eds.) *Social ontology and collective intentionality: Critical essays on the philosophy of Raimo tuomela with his responses* (3-18). Cham: Springer.

- Łukasiewicz, Jan (1970/1920) "On three-valued logic" in (O. Wojtasiewicz, Trans.) L. Borkowski (Ed.), *Selected works*. Amsterdam: North Holland Publishing Company.
- Lyons, David (1965) *Forms and limits of utilitarianism*. Oxford: Clarendon.
- Mackie, John L. (1965) "Causes and conditions". *American Philosophical Quarterly*, 2(4): 245-64.
- Mackie, John L. (1974) *The cement of the universe: A study of causation*. Oxford: Clarendon.
- Mayr, Erasmus (2011) *Understanding human agency*. Oxford: Oxford University Press.
- McDermott, Michael (1995) "Redundant causation". *British Journal for the Philosophy of Science*, 46(4): 523-44.
- McGrath, Sarah (2005) "Causation by omission: A dilemma". *Philosophical Studies*, 123(1/2): 125-48.
- McKenna, Michael (2012) *Conversation & responsibility*. New York: Oxford University Press.
- McLaughlin, James Angell (1925-26) "Proximate cause". *Harvard Law Review*, 39(2): 149-99.
- Mele, Alfred R. (1983) "Akrasia, reasons, and causes". *Philosophical Studies*, 44(3): 345-68.
- Mele, Alfred R. (1987) "Intentional action and wayward causal chains: The problem of tertiary waywardness". *Philosophical Studies*, 51(1): 55-60.
- Mele, Alfred R. (1992) *Springs of action: Understanding intentional behavior*. New York: Oxford University Press.
- Mele, Alfred R. (2003) "Agents' abilities". *Nous*, 37(3): 447-70.
- Menzies, Peter (1989) "Probabilistic causation and causal processes: A critique of Lewis". *Philosophy of Science*, 56(4): 642-63.
- Mill, John Stuart (2008/1859) "On liberty" in J. Gray (Ed.), *On liberty and other essays*. Oxford: Oxford University Press.
- Mill, John Stuart (2011/1843) *A system of logic, ratiocinative and inductive*. Cambridge: Cambridge University Press.
- Moore, G. E. (1912) *Ethics*. London: Oxford University Press.
- Moore, Michael S. (2009) *Causation and responsibility: An essay in law, morals, and metaphysics*. Oxford: Oxford University Press.
- Morton, Adam (1975) "Because he thought he had insulted him". *Journal of Philosophy*, 72(1): 5-15.
- Nagel, Thomas (1979) "Moral luck" in *Mortal questions*. Cambridge: Cambridge University Press.
- Nefsky, Julia (2012) "Consequentialism and the problem of collective harm: A reply to Kagan". *Philosophy and Public Affairs*, 39(4): 364-95.
- Nefsky, Julia (2015) "Fairness, participation, and the real problem of collective harm" in *Oxford studies in normative ethics* (Vol. 5: 245-71).
- Nefsky, Julia (2017) "How you can help, without making a difference". *Philosophical Studies*, 174(11): 2743-67.



- Nefsky, Julia (2019) "Collective harm and the inefficacy problem". *Philosophy Compass*, 14:e12587(4).
- Nefsky, Julia. (2021). *Participation, progress and superfluity*. Paper presented at the workshop *Small Acts, Big Harms*, Helsinki.
- Norcross, Alastair (1997) "Comparing harms: Headaches and human lives". *Philosophy & Public Affairs*, 26(2): 135-67.
- Norcross, Alastair (2004) "Puppies, pigs, and people: Eating meat and marginal cases". *Philosophical Perspectives*, 18(1): 229-45.
- Norcross, Alastair (2005) "Harming in context". *Philosophical Studies*, 123(1/2): 149-73.
- Nozick, Robert (1974) *Anarchy, state, and utopia*. New York: Basic Books.
- O'Neill, Onora (1985) "Between consenting adults". *Philosophy and Public Affairs*, 14(3): 252-77.
- O'Neill, Onora (1989) *Constructions of reason: Explorations of Kant's practical philosophy*. Cambridge: Cambridge University Press.
- Olson, Mancur (1965) *The logic of collective action: Public goods and the theory of groups*. Cambridge, Mass.: Harvard University Press.
- Palmer, T. N., A. Döring, & G. Seregin (2014) "The real butterfly effect". *Nonlinearity*, 27(9): R123-R41.
- Parfit, Derek (1984) *Reasons and persons*. Oxford: Clarendon Press.
- Parfit, Derek (1986) "Comments". *Ethics*, 96(4): 832-72.
- Parfit, Derek (1997) "Equality and priority". *Ratio*, 10(3): 202-21.
- Parfit, Derek (2011) *On what matters* (Vol. 1). Oxford: Oxford University Press.
- Paul, L. A. (1998) "Problems with late preemption". *Analysis*, 58(1): 48–53.
- Paul, L. A., & Edward J. Hall (2013) *Causation: A user's guide*. Oxford: Oxford University Press.
- Peels, Rik (2015) "A modal solution to the problem of moral luck". *American Philosophical Quarterly*, 52(1): 73-87.
- Penczek, Alan (1997) "Disjunctive properties and causal efficacy". *Philosophical Studies*, 86(2): 203-19.
- Petersson, Björn (2004) "The second mistake in moral mathematics is not about the worth of mere participation". *Utilitas*, 16(03): 288-315.
- Petersson, Björn (2013) "Co-responsibility and causal involvement". *Philosophia*, 41(3): 847-66.
- Petersson, Björn (2018) "Over-determined harms and harmless pluralities". *Ethical Theory and Moral Practice*, 21(4): 841-50.
- Petersson, Björn (2019) "Too many omissions, too much causation?" in T. Hansson Wahlberg & R. Stenwall (Eds.) *Maurinian truths: Essays in honour of Anna-Sofia Maurin on her 50th birthday*. Lund: Department of Philosophy, Lund University.
- Pinkert, Felix (2015) "What if I cannot make a difference (and know it)". *Ethics*, 125(4): 971-98.

- Portmore, Douglas W. (2018) "Teleological reasons" in D. Star (Ed.), *The Oxford handbook of reasons and normativity* (764-83). New York: Oxford University Press.
- Quinn, Warren S (1990) "The puzzle of the self-torturer". *Philosophical Studies*, 59(1): 79-90.
- Rabinowicz, Wlodek (1989) "Act-utilitarian prisoner's dilemmas". *Theoria*, 55(1): 1-44.
- Rabinowicz, Wlodek (2009) "Incommensurability and vagueness". *Proceedings of the Aristotelian Society*, 83: 71-94.
- Raffman, Diana (2000) "Is perceptual indiscriminability nontransitive?". *Philosophical Topics*, 28(1): 153-75.
- Ramachandran, Murali (1997) "A counterfactual analysis of causation". *Mind*, 106(422): 263-77.
- Rawls, John (1971) *A theory of justice*. Cambridge, Mass.: Harvard University Press.
- Rawls, John (1989) "Themes in Kant's moral philosophy" in E. Förster (Ed.), *Kant's transcendental deductions*. Stanford: Stanford University Press.
- Rolf, Bertil (1981) *Topics on vagueness* (Doctoral thesis). Lunds universitet.
- Rosen, Gideon (2015) "The alethic conception of moral responsibility" in R. Clarke, M. McKenna, & A. M. Smith (Eds.) *The nature of moral responsibility* (65-88). New York: Oxford University Press.
- Ross, William David (2002/1930) *The right and the good*. Oxford: Clarendon Press.
- Rousseau, Jean-Jacques (1984/1755) *Discourse on inequality* (M. Cranston, Trans.). Harmondsworth: Penguin Books.
- Russell, Bertrand (1912-13) "On the notion of cause". *Proceedings of the Aristotelian Society*, 7: 1-26.
- Russell, Bertrand (1923) "Vagueness". *Australasian Journal of Philosophy*, 1(2): 84-92.
- Salmon, Wesley C. (1984) *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.
- Salmon, Wesley C. (1994) "Causality without counterfactuals". *Philosophy of Science*, 61(2): 297-312.
- Sandberg, Joakim (2011) "'My emissions make no difference': Climate change and the argument from inconsequentialism". *Environmental Ethics*, 33(3): 229-48.
- Sandler, Ronald (2010) "Ethical theory and the problem of inconsequentialism: Why environmental ethicists should be virtue-oriented ethicists". *Journal of Agricultural and Environmental Ethics*, 23(1-2): 167.
- Sartorio, Carolina (2004) "How to be responsible for something without causing it". *Philosophical Perspectives*, 18(1): 315-36.
- Sartorio, Carolina (2006) "Disjunctive causes". *Journal of Philosophy*, 103(10): 521-38.
- Sartorio, Carolina. (2015). *Resultant luck and the thirsty traveler*. Retrieved from <https://sartorio.arizona.edu/papers/>. Originally published in *Method* 4(6), 153-171.
- Sartorio, Carolina (2016) *Causation and free will*. Oxford: Oxford University Press.
- Scanlon, Thomas (2008) *Moral dimensions: Permissibility, meaning, blame*. Cambridge, Mass.: Belknap Press of Harvard University Press.

- Schaffer, Jonathan (2000) "Causation by disconnection". *Philosophy of Science*, 67(2): 285-300.
- Schaffer, Jonathan (2001) "Causes as probability raisers of processes". *The Journal of Philosophy*, 98(2): 75-92.
- Schaffer, Jonathan (2005) "Contrastive causation". *Philosophical Review*, 114(3): 327-58.
- Schaffer, Jonathan (2012) "Causal contextualisms" in M. Blaauw (Ed.), *Contrastivism in philosophy: New perspectives*. New York: Routledge.
- Schopenhauer, Arthur (1969/1818-19) *The world as will and representation* (E. F. J. Payne, Trans. Vol. 1). New York: Dover Publications.
- Schroeder, Mark Andrew (2007) *Slaves of the passions*. Oxford: Oxford University Press.
- Schwartz, David T (2017) *Consuming choices: Ethics in a global consumer age* (2nd ed.). Lanham, Md.: Rowman & Littlefield.
- Schwenkenbecher, Anne (2014) "Is there an obligation to reduce one's individual carbon footprint?". *Critical Review of International Social and Political Philosophy*, 17(2): 168-88.
- Searle, John (1990) "Collective intentions and actions" in J. M. P. Cohen, and M.E. Pollack (Ed.), *Intentions in communication*. Cambridge, Mass.: Bradford Books, MIT Press.
- Shoemaker, David (2015) *Responsibility from the margins*. Oxford: Oxford University Press.
- Shrader-Frechette, Kristin (1987) "Parfit and mistakes in moral mathematics". *Ethics*, 98(1): 50-60.
- Singer, Peter (1980) "Utilitarianism and vegetarianism". *Philosophy & Public Affairs*, 9(4): 325-37.
- Sinnott-Armstrong, Walter (2005) "It's not my fault: Global warming and individual moral obligations" in W. Sinnott-Armstrong & R. Howarth (Eds.) *Perspectives on climate change* (221–53). Amsterdam: Elsevier.
- Sinnott-Armstrong, Walter (2008) "A contrastivist manifesto". *Social Epistemology*, 22(3): 257–70.
- Skorupski, John (2010) *The domain of reasons*. Oxford: Oxford University Press.
- Skyrms, Brian (2004) *The stag hunt and the evolution of social structure*. New York: Cambridge University Press.
- Slote, Michael (2001) *Morals from motives*. Oxford: Oxford University Press.
- Smith, Angela (2005) "Responsibility for attitudes: Activity and passivity in mental life". *Ethics*, 115(2): 236-71.
- Smith, Angela (2008) "Control, responsibility, and moral assessment". *Philosophical Studies*, 138(3): 367-92.
- Smith, Angela (2015) "Attitudes, tracing, and control". *Journal of Applied Philosophy*, 32(2): 115-32.
- Snedegar, Justin (2017) *Contrastive reasons*. Oxford: Oxford University Press.
- Sorensen, Roy A. (1988) *Blindspots*. Oxford: Clarendon Press.

- Spiekermann, Kai (2014) "Small impacts and imperceptible effects: Causing harm with others". *Midwest Studies In Philosophy*, 38(1): 75-90.
- Stapleton, Jane (2008) "Choosing what we mean by causation in the law". *Missouri Law Review*, 73(2): 433-480.
- Strawson, P.F. (2008/1962) "Freedom and resentment " in *Freedom and resentment and other essays*. London: Routledge.
- Streumer, Bart (2007) "Reasons and impossibility". *Philosophical Studies*, 136(3): 351-84.
- Talbert, Matthew (2012) "Moral competence, moral blame, and protest". *The Journal of Ethics*, 16(1): 89-109.
- Talbert, Matthew (2015) "Responsibility without causation, luck, and dying of thirst: A reply to Sartorio". *Method*, 4: 283-96.
- Talbert, Matthew (2017) "Omission and attribution error" in D. K. Nelkin & S. C. Rickless (Eds.) *The ethics and law of omissions*. New York: Oxford University Press.
- Talbert, Matthew (2019) "The attributionist approach to moral luck". *Midwest Studies In Philosophy*, 43(1): 24-41.
- Talbert, Matthew (forthcoming) "Attributionist theories of moral responsibility" in D. K. Nelkin & D. Pereboom (Eds.) *The Oxford handbook of moral responsibility*. Oxford: Oxford University Press.
- Tappenden, Jamie (1993) "The liar and sorites paradoxes: Toward a unified treatment". *Journal of Philosophy*, 90(11): 551-77.
- Temkin, Larry S. (2000) "Equality, priority, and the levelling down objection" in M. Clayton & A. Williams (Eds.) *The ideal of equality*. Basingstoke: Macmillan.
- Temkin, Larry S. (2012) *Rethinking the good: Moral ideals and the nature of practical reasoning*. New York: Oxford University Press.
- Thalberg, Irving (1984) "Do our intentions cause our intentional actions?". *American Philosophical Quarterly*, 21(3): 249-60.
- Thomson, Judith Jarvis (1976) "Killing, letting die, and the trolley problem". *The Monist*, 59(2): 204-17.
- Thomson, Judith Jarvis (2008) "Some reflections on Hart and Honoré, causation in the law 1" in M. H. Kramer, C. Grant, B. Colburn, & A. Hatzistavrou (Eds.) *The legacy of H.L.A. Hart: Legal, political, and moral philosophy*. Oxford: Oxford University Press.
- Touborg, Caroline (2017) "Hasteners and delayers: Why rains don't cause fires". *Philosophical Studies*, 175: 1557-76.
- Touborg, Caroline (2018) *The dual nature of causation: Two necessary and jointly sufficient conditions* (Doctoral thesis). University of St Andrews.
- Tye, Michael (1990) "Vague objects". *Mind*, 99(396): 535-57.
- Tye, Michael (1994) "Sorites paradoxes and the semantics of vagueness". *Philosophical Perspectives*, 8: 189-206.
- Unger, Peter (1979) "There are no ordinary things". *Synthese*, 41(2): 117-54.
- van de Poel, Ibo (2011) "The relation between forward-looking and backward-looking responsibility" in A. N. Vincent, I. van de Poel, & J. Hoven (Eds.) *Moral responsibility: Beyond free will and determinism* (37-52). Dordrecht: Springer.

- Van Inwagen, Peter (1978) "Ability and responsibility". *The Philosophical Review*, 87(2): 201-24.
- Van Inwagen, Peter (1983) *An essay on free will*. Oxford: Clarendon Press.
- Wallace, R. Jay (1994) *Responsibility and the moral sentiments*. Cambridge, Mass.: Harvard University Press.
- Watson, Gary (1987) "Responsibility and the limits of evil: Variations on a strawsonian theme" in F. Schoeman (Ed.), *Responsibility, character, and the emotions: New essays in moral psychology*. Cambridge: Cambridge University Press.
- Watson, Gary (2004/1975) "Free agency" in G. Watson (Ed.), *Agency and answerability: Selected essays*. Oxford: Clarendon.
- Werkmäster, Jakob Green (2019) *Reasons and normativity* (Doctoral thesis). Lund University.
- Wieland, Jan Willem, & Philip Robichaud (2017) *Responsibility: The epistemic condition*. Oxford: Oxford University Press.
- Wieland, Jan Willem, & Rutger van Oeveren (2020) "Participation and superfluity". *Journal of Moral Philosophy*, 17(2): 163-87.
- Williams, Bernard (1981) *Moral luck*. Cambridge: Cambridge University Press.
- Williamson, Timothy (1994) *Vagueness*. London: Routledge.
- Wolf, Susan (1990) *Freedom within reason*. New York: Oxford University Press.
- Wood, Allen W. (1999) *Kant's ethical thought*. Cambridge: Cambridge University Press.
- Woodward, James (2003) *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.
- Woollard, Fiona (2015) *Doing and allowing harm*. Oxford: Oxford University Press.
- Wright, Richard W. (1985) "Causation in tort law". *California Law Review*, 73(6): 1735-828.
- Wright, Richard W. (1988) "Causation, responsibility, risk, probability, naked statistics, and proof: Pruning the bramble bush by clarifying the concepts". *Iowa L. Rev.*, 73: 1001-77.
- Wright, Richard W. (2013) "The NESS account of natural causation: A response to criticisms" in M. Stepanians & B. Kahmen (Eds.) *Critical essays on "causation and responsibility"* (13-66). Berlin: De Gruyter.
- Wringe, Bill (2016) "Collective obligations: Their existence, their explanatory power, and their supervenience on the obligations of individuals". *European Journal of Philosophy*, 24(2): 472-97.
- Young, Iris Marion (2011) *Responsibility for justice*. New York: Oxford University Press.
- Zagzebski, Linda Trinkaus (2004) *Divine motivation theory*. Cambridge: Cambridge University Press.
- Zimmerman, Michael J. (2002) "Taking luck seriously". *Journal of Philosophy*, 99(11): 553-76.
- Zimmerman, Michael J. (2011) *The immorality of punishment*. Peterborough, Ont.: Broadview Press.



## Reasons, Blame, and Collective Harms

---

Do you have a climate-change-related reason to refrain from going for a leisure drive in a gas-guzzling car? And do you have a reason not to take a shortcut across a beautiful lawn in case the lawn will be ruined if enough people cross it? Questions like these are not easy to answer. On the one hand, it seems that you have such reasons since acting in the relevant way contributes to a bad outcome. On the other, it seems that you lack such reasons since your particular act makes no difference to the outcome. Climate change will occur just the same whether you go for a leisure drive or not, and the lawn will look just the same whether you cross it or not.

Building on contemporary accounts of causation, this book suggests an account of outcome-related reasons that can explain both kinds of intuitions. The different kinds of intuitions stem from different perspectives. It is argued that the first perspective according to which you have an outcome-related reason to act in the relevant way often is the more correct one. In addition to giving intuitively correct verdicts about what reasons you have in collective harm cases like these, the suggested account can explain our intuitions about reasons in pre-emption cases, overdetermination cases, switching cases, omission cases, Frankfurt-style cases, the difficult case of the thirsty traveller, and more.

Besides giving an account of outcome-related reasons, this book also gives a corresponding account of when you are blameworthy for an action, omission or outcome. In doing so, it connects three different debates, the one on collective harms, the one on causation, and the one on moral responsibility, and does so in new and illuminating ways.



LUND  
UNIVERSITY

Department of Philosophy  
The Joint Faculties of Humanities and Theology  
Lund University

ISBN 978-91-89213-95-1

