The London School of Economics and Political Science

The Doing/Allowing Distinction:

Causal Relevance and Moral Significance

Camilla Francesca Colombo

A thesis submitted to the Department Philosophy, Logic and Scientific Method of the London School of Economics for the degree of Doctor of Philosophy,
London, 1st October 2018

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others.

I confirm that material from Chapter 4 of this thesis will be published as "Doing, Allowing, Gains, and Losses" in *Ethical Theory and Moral Practice* (forthcoming).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 63,570 words.

Camilla Francesca Colombo

Statement of use of third party for editorial help

I can confirm that my thesis was copy edited for conventions of language, spelling and grammar by Adele Jourdan and Sarah Taylor.

Abstract

Intuitively, an agent who *does* harm behaves differently from an agent who *allows* harm to happen. This thesis examines the distinction between doing harm and merely allowing it to occur. I argue that this distinction is morally relevant, and doing harm is harder to justify than allowing harm, but that there is not always a fact of the matter how the distinction ought to be drawn.

In Chapters 1 and 2, I survey the main alternative accounts for explaining the difference between “doing” and “allowing”. I compare causal approaches, which distinguish doing and allowing on the basis of how an agent caused an outcome, with “norm-based” accounts, which explain the distinction appealing to independent moral features. I conclude that a “mixed” causal account, such as Hitchcock's self-contained network model, is the most promising for tracking doing/allowing classifications.

I then examine whether this distinction is morally relevant. I outline two theoretical hypotheses, the “positive” and the “negative” theses. The former argues that there is a fact of the matter whether an action is “doing” or “allowing”, and this classification is morally significant; the latter that there might or might not be such a fact of the matter, but in any case this distinction is not morally relevant. In Chapter 3, I critique an influential strategy for settling this issue, that is, comparing “fully-equalized cases”. In Chapter 4, I consider the import of “framing effects”. Despite attempts to use fully-equalized cases or evidence of framing in support of either thesis, these strategies are not compelling.

In Chapters 5 and 6, I present my alternative thesis, which relies on the self-contained network model. I define “doings” as instances where an outcome counterfactually depends on the agent, within a “self-contained” network, and “allowings” as instances where the outcome depends on the agent, within a “non-self-contained” network. This classification captures whether an agent is causally relevant to an outcome in a specific way; nonetheless, the identification of “self-contained” networks incorporates agents' empirical and normative expectations. The distinction is thus morally relevant, as it (also) captures moral considerations, but may be ultimately ambiguous, as there may not always be a fact of the matter as to how the distinction should be drawn.

List of Contents

List of Figures and Tables

Acknowledgments

First and foremost, I am grateful to my supervisor Katie Steele. Katie provided extensive and extremely helpful feedback on countless drafts of my thesis, no matter how busy she was. Katie paid care and attention to my work, and her advice has shaped and improved this thesis greatly. Her support during these four years has been invaluable.

I am also grateful to Susanne Burri for being a fantastic second supervisor. Especially in my last year, Susanne has provided insightful comments, and even more helpful advice and support. Susanne is both kind and efficient, and she always had time when I needed to discuss an issue.

I thank the LSE Philosophy Department for my four-year scholarship, and for being an outstanding work environment. I learned a lot from my fellow PhD students, and I should offer special thanks to Kamilla Buchter, Aron Vallinder and Christina Easton.

Finally, I would like to thank Christian List and Fiona Woollard for being my examiners. Their thorough comments and feedback have helped me perfectioning this thesis, and have given me a lot to think about for future developments.

I thank my brother and my sister for acting as my personal debate club, and Nicole, my brilliant friend.

Introduction

The idea that *doing* and *allowing* amount to two distinct forms of human behaviour, with different meaning and significance, strikes us as intuitive and reasonable. There are reasons to believe that there is a fundamental asymmetry between "doing" and "allowing" in general; in this thesis, however, I focus narrowly on the dichotomy between doing *harm* and allowing *harm* to occur, as opposed to doing good or allowing good to occur, since the former has been more thoroughly investigated in moral philosophy. We do, in fact, use this distinction in everyday life; when forming moral judgements, at least prior to reflection, we seem to share an overwhelming intuition that doing harm is somehow worse, or harder to justify, than allowing it to occur. This thesis sets out to examine this *doing/allowing distinction.*

The doing/allowing debate in moral philosophy revolves around two main questions: i) where to draw the line between doings and allowings, and ii) whether this distinction matters morally. That is, whether doing behaviours are descriptively different from allowing behaviours, and whether doing harm is harder to justify than allowing harm. In this thesis, I strive to keep these two questions apart, and to address the descriptive and the normative task separately. As I examine these issues, nonetheless, I observe that the most promising descriptive accounts of the doing/allowing distinction incorporate moral considerations. At the same time, disputes about whether the distinction is morally significant turn out to be intimately tied up with disputes about how to distinguish doing from allowing in the first place. I conclude that disagreement regarding doing and allowing classifications and the relative moral significance of such behaviours depends on both empirical and normative features. My proposed account of the doing/allowing distinction ultimately aims to explain persistent disagreement in everyday use of this dichotomy and in moral theorising, while preserving the intuition that doing/allowing classifications capture morally relevant features of behaviours so described.

In this sense, this thesis aims to bring together two different lines of investigation regarding the doing/allowing distinction, which do not often engage with each other. In the first camp, authors like Philippa Foot, Warren Quinn, Jeff McMahan, Frances Kamm, and Fiona Woollard strive to explain the different meaning that commonsense morality seems to attach to "doings" and "allowings". They take seriously our intuitive judgements about specific cases, and try to build upon them a systematic account of the doing/allowing distinction, which justifies the insight that "doing is worse than allowing". Still in this first camp, authors like Jonathan Bennett, Shelly Kagan, and James Rachels have however challenged the idea that the doing/allowing distinction, in spite of its central role in everyday moral practice, is morally relevant. They argue that the different significance we attach to doings and allowings is not justified after all, either because this distinction is grounded in morally irrelevant features or because it disappears upon careful analysis. In the second camp, authors like Tamara Horowitz, Walter Sinnott-Armstrong, and Fiery Cushman have been investigating the doing/allowing distinction as a cognitive bias or, more charitably, as the byproduct of our (flawed) moral reasoning skills. The role of our intuitions about specific cases is thus downplayed, as moral intuitions seem, under closer scrutiny, generally unreliable, controversial and frame-dependent.

In this thesis, I take from the first camp the idea that, if we aim to explain commonsense morality, we need to account for our use of the doing/allowing distinction. This means that, to some extent, we cannot easily dismiss the intuitive judgement that doing harm, all other things being equal, is worse than allowing harm, and that these two conducts are somehow distinct. Nonetheless, I also look at evidence of disagreement and framing effects. From the second camp I thus take the idea that our moral intuitions, if not unreliable, might be context- and agent-dependent. My project aims to preserve both insights.

From this intermediate position, I thus argue that doing/allowing classifications are morally significant, as "composite judgements" which incorporate considerations that matter for moral evaluation. As such, these descriptions of behaviours should be taken seriously and can legitimately serve our everyday moral practice. On the other hand, doing/allowing classifications are also less stable, more controversial and less clear-cut than some might hope. Different agents, contexts and framings may make salient different considerations, and deliver different doing/allowing descriptions. Unlike cognitive bias theorists, nonetheless, we need not conclude that our moral intuitions lead us completely astray. In many cases, doing/allowing classifications are "robust" and agreed-upon; in these circumstances, we should keep our intuitive judgement that doing harm is morally worse than allowing harm, all other things being equal. In controversial cases, which I argue are often under-described, I suggest that we should look at those things which are *not* "equal", that is, are frame- or agent-dependent.

Summary of chapters

In Chapter 1 and Chapter 2, I analyse how we should best conceptualise the doing/allowing distinction, and I survey the main alternative accounts for explaining the difference between doing and allowing, setting aside the problem whether this distinction is morally significant. Specifically, I compare causal approaches, which distinguish doing and allowing on the basis of how an agent caused an outcome, with what I call "norm-based" accounts, which attempt to explain the distinction by appealing to independent moral features. I conclude that a "mixed" causal account, like Christopher Hitchcock's self-contained network model, which takes on board norm-based considerations in the identification of specific types of causal relations, amounts to the most suitable tool for tracking our everyday use of the doing/allowing distinction.

I then start examining the question of whether the doing/allowing distinction is morally relevant, and I outline two main theoretical hypotheses to this end, which I call the "negative" and the "positive" theses. The positive thesis amounts to the position that there is a fact of the matter whether an action is an instance of "doing" or else "allowing", and this classification is morally significant *per se*, that is, it is a fundamental one and cannot be expressed by other moral features or principles. The negative thesis, on the other hand, argues that there might or might not be a fact of the matter whether an action is a "doing" or an "allowing", but in any case this distinction is not morally relevant *per se*. The middle chapters of this thesis survey different arguments and evidence in favour of both the positive or the negative thesis.

In Chapter 3, I discuss and critique an influential strategy employed in the literature for settling this issue, which amounts to describing "fully-equalized cases", so-named because they are designed to test agents' responses to the doing/allowing distinction alone. In Chapter 4, I consider the significance of "framing effects", that is, the fact that agents' classification of the (seemingly) same action as an instance of "doing" or else "allowing" can sometimes depend on the precise description or "framing" of the action. I conclude that, despite the attempts to use fully-equalized cases or evidence of framing effects in support of either the positive or the negative thesis, neither of these strategies provides a compelling case in favour of either position.

More specifically, persistent disagreement and lack of consensus, in both people's intuitions and moral theorising, seem to amount to the most prominent feature of the comparisons between allegedly fully-equalized cases. The analysis of the nature and source of such disagreement suggests that the project of disentangling the doing/allowing distinction from any other relevant aspect of the context, while *prima facie* promising, is ultimately unsuccessful. In any case, what should be concluded from the analysis of fully-equalized cases remains at least controversial.

Supporters of the negative thesis argue that framing effects challenge the moral significance of the distinction. If our intuitions about what counts as "doing" and what counts as "allowing" are influenced by seemingly morally irrelevant features, such as the description one uses or the order information is presented, then we have reasons to reject the idea that "doing" should be harder to justify than "allowing". I discuss in particular the "Asian flu" case, which has been employed to show that the doing/allowing distinction collapses into reasoning biases and psychological attitudes to risk. Nonetheless, it is open to supporters of the positive thesis to argue that while framing effects are real, they have no implications for the significance of the doing/allowing distinction. Rather, they point to flaws in our moral reasoning that need to be, and can be, corrected on reflection. In this way, framing effects do not necessarily amount to evidence in favour of either thesis, though they do suggest that the distinction might be ultimately ambiguous. With "ambiguous" I mean here that there may not always be a "correct" classification of actions as doing or allowings, as in some cases different classification might be tenable, depending on the agent and of the framing of the case.

It thus looks like neither the negative nor the positive theses emerge as a clear winner. Beyond that, both theses seem to account for aspects we intuitively feel are right: the positive thesis captures the intuition that, with doing/allowing classifications, we are onto something morally significant, but also requires us to downplay evidence of disagreement. The negative thesis, conversely, may adequately explain disagreement, controversial cases and examples of framing, but may go against strong intuitions of moral relevance.

In Chapter 5, I suggest that a different proposal, which I call the "alternative thesis" may deliver a way out of this standoff, and preserve both aspects. The alternative thesis takes the following form: the doing/allowing distinction may be ultimately ambiguous, that is, it can be frame- and agent-dependent, but it is nonetheless morally relevant, that is, *all other things being equal*, "doing" is harder to justify than "allowing". To do so, the alternative thesis must both argue that frame-dependency does not rule out moral significance (*contra* the positive and the negative thesis) and that the doing/allowing distinction is not morally relevant *per se* (*contra* the positive thesis), but still captures morally relevant features. Specifically, I define the alternative thesis as a "reductionist" proposal, insofar as it does not consider the doing/allowing distinction as a fundamental feature of actions. I further argue that the attractiveness of the alternative thesis lies in its explanatory advantages, which I outline in terms of three desiderata: i) it captures the idea that "doing" and "allowing" are both causing in a specific way, ii) it explains and allows for disagreement and frame-dependency in doing/allowing classifications, and iii) it accounts for the fact that the distinction is morally relevant.

My idea is to cash out the alternative thesis using Hitchcock's self-contained network model: I define "doing" actions as instances where an outcome counterfactually depends on the agent, within a "self-contained" causal network. "Allowing" actions, on the other hand, describe situations where the outcome counterfactually depends on the agent, within a "non-self-contained" causal network. Doing/allowing classifications, within this model, thus clearly track whether an agent is causally relevant to an outcome in an "act" (doing) rather than in an “omission” way (allowing).

The identification of “self-contained” causal networks, however, depends in this model upon which values are assigned to the variables, and, specifically, which value is set as the “default” for all the variables in the network. This feature of the model reflects the insight that doing and allowing are defined with reference to the “normal” course of events. The assignment of default values thus incorporates agents' expectations and judgements regarding both descriptive and normative features of the context. So doing/allowing classifications may vary depending on what agents think will happen or should happen, and depending on the specific framing people may infer different “normal” courses of events.

In most straightforward, detailed and agreed-upon cases, doing/allowing classifications reliably track other morally relevant considerations such as whether the agent intended the harm or the agent acted violating a standard norm. I show how the self-contained network model can incorporate these features in Chapter 6. In particular, I argue that doing/allowing classifications may be best interpreted as a “composite judgement”, which tracks different moral and empirical considerations. On the other hand, when cases are unfamiliar, under-described, or pitch different norms against one another in a fairly extreme way, different doing/allowing classifications are reasonable and justifiable, as different default values are legitimate. In these cases, disagreement is to be expected.

This model, I argue, satisfies all three desiderata. On my proposed account, the doing/allowing distinction clearly has moral significance, as it captures morally relevant considerations in a composite judgment. It cannot, however, ultimately be reduced to a single moral principle which consistently explains all the doing/allowing distinctions we draw and why we disagree about them. However, my model enables us to analyse doing/allowing attributions on a case-by-case basis.

1. Analysis of Doing *vs* Allowing: Causal Accounts

The idea that *doing harm* and *allowing a harm* to occur amount to two distinct forms of conduct, with different significance and meaning, strikes us as intuitive and reasonable. We do, in fact, use this distinction in real life for many practical circumstances, such as assigning blame and responsibility and calculating compensations. When forming moral judgements, specifically, at least prior to reflection, we seem to share an overwhelming intuition that doing harm is somehow worse than allowing a harm to occur, and should rank higher in terms of the magnitude of the wrongdoing.

While the moral significance and practical implications of the doing/allowing harm distinction are my ultimate interest, I set these concerns aside in the first and second chapters of this thesis. For the moment, my task will be an *analysis* of this commonly perceived fundamental dichotomy in the realm of human behaviour, the doing/allowing distinction, which is often employed in non-moral cases as well. This analysis will take the form of a search for the most suitable and convincing explanation for the fact that, to use a paradigmatic pair of cases, when I throw a rock at a window, I am “doing” something, while when I do not check and fix my boiler, I am “allowing” something to occur.

1.1 Doing as Causing

One of the most widespread insights in the doing/allowing literature is that this distinction has something to do with *causation*. Intuitively, when I throw the rock at the window I cause the window to break, and, possibly, a subsequent threat to other people; on the other hand, my negligence in checking the status of my appliances does not cause – or at least does not cause *in the same way* – the boiler to break, with the related hazard. While this idea appears sound and persuasive, the question of which account of causation is more suitable for cashing out the doing/allowing distinction is by no means a straightforward one to answer.

The theoretical approaches to the concept of causation are extremely numerous and varied, and it is not my place here to provide a detailed review and a comparative assessment of all alternative frameworks. In order to navigate my way across the huge literature on the nature of causal connection, I focus narrowly on how well different causal accounts fit the way we talk about human agency in particular. While all causal models can, at least in principle, be applied to human agency, my aim here is to specify desiderata that can then be used to assess how well a causal account tracks and makes sense of our employment of the doing/allowing distinction. I will articulate these desiderata as I go along my survey.

The alternatives on the table when it comes to the nature of causal relations can be classified into two main families, which may be referred to as *process theories* and *difference-making accounts* respectively. The general idea behind process theories is that causation amounts to some concrete physical connection, and causing is producing. Difference-making accounts, on the other hand, rely on the insight that causing is making more likely to occur.

In the present chapter, I will address both difference-making and process theories of causation, as well as more informal versions of process theories. I will start from difference-making accounts, and then move on to process theories. I argue that both accounts, at least in their philosophy of science formulation, have some key gaps when it comes to distinguishing doing and allowing in human agency. In both cases, moral philosophers have attempted to fill these gaps. I will survey some of these attempts, and conclude that refined difference-making accounts are most promising in this respect. Some missing details of these refined difference-making accounts will be spelled out in Chapter 2.

1.2 Difference-making accounts

The basic intuition behind difference-making accounts is that causal relations are defined in terms of counterfactual conditionals. Roughly, as Lewis (1986) puts it, we can capture the notion of causal relation with the following definition: an event *E* causes an event *F* if, had *E* not occurred, *F* wouldn't have occurred either. With respect to the doing/allowing distinction, we can thus make sense of this dichotomy as tracking the *impact* of an agent's conduct on the upshot: doing *x* would mean causing *x*, in the sense that had one not acted in that way, *x* would not have occurred. Allowing *x* would then be not causing *x*, in the sense that had one not acted in that way, *x* would have occurred anyway.

As mentioned above, difference-making accounts of causation share the idea that causing is making an upshot more likely. The relation between an agent and the upshot, therefore, may be also a probabilistic one, thus taking the counterfactual form "*E* causes *F* just in case if *E* had not occurred, *F* would have been less likely". The counterfactual approach, in its general probabilistic form, is implemented in slightly different ways by various authors, including von Wright (1975), Woodward (2004), Pearl (2000), Price and Menzies (1993). These differences do not concern us here. As mentioned, in this section, I focus rather on the general merits and problems of counterfactual accounts.

Let's start testing the counterfactual approach to causation with a couple of classical doing/allowing examples. When I break the window by throwing a rock, I cause the broken window, as reflected by the truth of the following counterfactual: if event *E*, my throwing the rock, had not happened, in most counterfactual scenarios event *F*, the window breaking, wouldn't have occurred either, or, at least, it would have been less likely. On the other hand, when I stand by while a rock slips, crushing into someone beneath, I do not cause the injury, as reflected by the falsity of the following counterfactual: event *F* (the rock crushing a person) would have occurred (or, at least, would have likely occurred) in most counterfactual situations where action *E*, my standing aside, had not occurred; precisely, *F* would not have occurred only in the specific scenario where I had tried to stop the rock and my efforts had been successful. Recall now that causing means doing and not causing means allowing. So the analysis seems to fit our intuitions: I “did” something throwing the rock, and I “allowed” something to happen when I did not stop the slipping rock.

I will now briefly address a salient worry about difference-making accounts, which is preemption; I will then move to what I consider the key issue, and first desideratum – distinguishing between acts and omissions.

1.2.1 Preemption

Since Lewis' first analysis of causation in terms of counterfactuals, a wide number of counterexamples and objections have been put forward in the literature, resulting to several revisions and refinements of the simple counterfactual framework sketched above, and which are common to most major contemporary accounts. Peter Menzies (2003) classified the main objections as *context-sensitivity*, *temporal asymmetry*, *preemption* and *transitivity*.[1] It is beyond the scope of this chapter to address all of these issues, so I focus here on the most relevant to the doing/allowing distinction. In this section, I briefly discuss preemption, while in 1.4.1 I deal with a specific instance of context-dependency.

Take the following simple example of preemption:[2]

An assassin, Alice, poisons her victim's drink. To be sure of the success of her mission,

[1] Menzies (2003), pp. 5–13.

[2] I take this formulation of the Backup example from Hitchcock (2009), p. 588.

> however, she also has a Backup, Bob, who is ready to poison the drink if Alice fails in her attempt.

This apparently simple case provides a powerful and notorious counterexample to standard counterfactual analysis: intuitively, we consider Alice's poisoning as causing the death of the victim. However, the existence of a genuine causal connection is ruled out by counterfactual analysis: due to the presence of the backup (Bob) the victim would have died anyway, had Alice not poisoned her drink.

This bug within counterfactual accounts looks like bad news if we want to ground doing/allowing classifications in this causation model. If "doing" amounts to causing, indeed, we could not define preemption cases as instances of doing, which is clearly unacceptable if we are interested in keeping track of our intuitions and practical use of the distinction. It seems untenable, indeed, to argue that Alice "did not" kill the victim, merely because there was a backup who could have taken over had she failed. An adequate counterfactual account, therefore, would need to address the issue of preemption, so as to make sense of the way we understand human agency.

Different approaches have been put forward in the literature to solve the puzzle preemption cases pose to counterfactual analysis. Some authors, like Lewis (1986), Paul (2000) and Coady (2004), suggest that a more fine-grained description of events might be the solution: the upshot, in Backup, could be defined as "the victim dies at t_1, drinking that kind and specific amount of poison". In this sense, when the outcome is characterised in more precise terms, we get the result that only Alice, and not Bob, could have caused it as, say, Bob would have poured the poison later or used a different amount of poison. Sartorio (2005) proposes a different approach, which is to argue that, in Backup, we can still identify a group, formed in this case by Bob and Alice, which counterfactually causes the upshot. While this solution does not allow us to claim that Alice "did" kill the victim, we can at least conclude that there was a "doing" conduct involved, and that Alice was part of the group who caused the upshot to occur. While the first solution sounds convincing, it also seems that adding precision would require us to describe the outcomes in details which are arguably irrelevant, such as the precise time the upshot occurred. This, however, does not look promising if we aim to capture our intuitions about human agency.

More complex preemption examples, moreover, seem even harder to accommodate;[3] there is thus huge disagreement in the literature as to whether any proposal can successfully account for all instances of preemption. Some authors, like Hitchcock (2001), have therefore suggested that counterfactual theory might as well bite the bullet and simply provide some useful tools for

[3] See, for instance, the "Spell case" in Schaffer (2000), p. 165.

describing causal relations among agents and upshots in preemption cases. For the time being, I note that counterfactual accounts may have difficulty in making sense of our intuitive assessments of causation in some "gimmicky examples"[4] like preemption. These tricky cases seem also to reveal that the counterfactual account of causation must be supplemented with something further that is sensitive to context in order to account for some problematic examples. This could mean that refined counterfactual accounts might be more "nuanced" and less elegant than one might have liked. I will come back to this point at the end of Chapter 2, when defending Hitchcock's model as a suitable account for capturing the doing/allowing distinction.

1.2.2 Distinguishing acts and omissions

Let's now leave technical difficulties like preemption aside and go back to the correct identification of doing and allowing conducts. What seems particularly appealing and convincing about counterfactual analysis is the idea of evaluating whether an agent amounts to a *difference maker* with respect to the outcome. As Kagan (1989) puts it,[5] this captures the insight that I cause an outcome when I interfere with the natural course of events, while I do not cause it when I do not interfere, as my existence would not make any difference to this natural course. The doing/allowing distinction, in this sense, would be tracking this crucial notion of interference. While the idea that "doing" behaviours amount to "causing as interfering" appears particularly straightforward, the understanding of "allowing" within the counterfactual model sketched above nonetheless seems unsatisfactory upon closer reflection. On the one hand, if allowing is "not causing", it remains unclear what should distinguish "allowing" behaviours from all other instances of non-causal relation between two events. In the *boiler* or in the *slipping rock* cases, as well as in classical "refraining" or "not aiding" examples, the characterisation of the agent's behaviour as "allowing" seems arguably to appeal to the fact that her conduct was, somehow, *causally relevant* to the upshot. This particular relation between the agent and the outcome may be different from the one between my throwing the rock and the window breaking, but, surely, it is also distinct from the connection between my brushing my teeth and the water boiling in the kitchen. Instances of allowing, in short, seem to refer to cases where the absence of the agent, or her inaction, are causally relevant to the outcome – like my *not* checking the boiler. Allowing cases are typically referred to as (causally relevant) *omissions*. If we want to make sense of the way we use "allowing" attributions, therefore, we need a causal

[4] Kagan (1989) uses this term to refer to cases which seem particularly tricky and not easy to accommodate within standard models of the doing/allowing distinction.

[5] Kagan (1989), pp. 92–101.

account where omissions are not dismissed and equated to all other causally unrelated events.

One might respond that counterfactual accounts do in fact treat omissions as causally relevant, as they can be conceived as difference makers, raising the probability of an outcome; in some sense, had I not neglected my boiler, it would have been less likely for it to break. Upon specific formulations and interpretations of counterfactuals, like using fine-grained descriptions of events, most omissions can be thus conceived of as difference makers. But in this case we have the opposite problem: we cannot distinguish omissions from other causes, and so it looks like behaviours that are intuitively allowings would be classified as doings. When talking about my negligence in checking the boiler, which seemingly classifies as an instance of allowing, however, we want both to maintain that my behaviour had an impact on the outcome, but also that I "did not" break the boiler.

While it looks like counterfactual analysis is going to be too crude to handle the doing/allowing distinction, it may be that it is simply underspecified (as already suggested above in relation to preemption), and needs to be supplemented with further detail. That is indeed what I will propose in the next chapter. For the time being, I note that this discussion has so far revealed two key desiderata for an account of causation which can capture the doing/allowing distinction. Firstly, we need to account for omissions as genuine causes. Secondly, cashing out the doing/allowing distinction further requires that we properly distinguish between two kinds of causally relevant behaviours, and between the different impact doing and allowing conducts seem to have on an upshot. Roughly, causal accounts of the doing/allowing distinction must allow for both *acts* and *omissions* to count as causally relevant, and provide a criterion for discriminating the former from the latter. To be clear, I do not take the act/omission distinction to be equivalent to the doing/allowing distinction;[6] as I will articulate more thoroughly in this thesis, I think that the latter is more complex and articulated than the former. Nonetheless, to make sense of the doing/allowing distinction, we still need to distinguish between two fundamental ways in which an agent can be causally relevant to an outcome, which I identify as the act/omission distinction. In the first two chapters of this thesis, I will thus use the two distinctions interchangeably.[7]

I will soon return (in 1.4) to the question of whether more sophisticated counterfactual accounts

[6] For instance, I think it is possible to allow harm by acting, like in the Impoverished Village example discussed in section 6.4 of this thesis.

[7] Specifically, I use the act/omission distinction to talk about the causal contribution of an agent to an outcome. As I will explain in Chapters 5 and 6, however, my account of the doing/allowing distinction allows for other considerations, beyond causal relevance, to be incorporated into these classifications. In this respect, the doing/allowing distinction does not coincide with the act/omission distinction. In "non moralised" examples, and in my analysis in Chapters 1 and 2, we can nonetheless talk of these two distinctions as equivalent, insofar as we are focusing on capturing the specific causal impact of an agent on an outcome.

can accommodate the act/omission distinction. But first, I examine whether traditional process theories can handle the distinction in a more straightforward way. I will argue that they cannot, despite looking promising in this regard (which is why the remainder of the chapter will dwell, for the most part, on sophisticated counterfactual approaches).

1.3 Process Theories

As mentioned above, the underlying intuition of process theories is that causation amounts to causal connections or chains of events involving some form of continuous change (Ducasse 1926), energy flow or transfer of energy (Skyrms 1984), a physical process (Salmon 1984, Mackie 1974) or some kind of transference of properties (Aronson 1971, Kistler 1998). When it comes to evaluating the role of a human agent or of a specific human conduct within one of these models, what we will be looking at is the agent's *contribution* to the outcome in terms of energy transfer or physical process.

Let's now move to how process theories can account for the doing/allowing distinction, by taking a fairly generic version of a physical connection model, like that proposed by Aronson (1971).[8] According to Aronson, causation is the transference of a specific quantity, such as, for instance, velocity, momentum, kinetic energy, or heat [9] by the means of the contact between cause and effect. Specifically, "A in 'A causes B' refers to an object that successfully transfers one of its quantities to the effect object."[10] We can then judge the role of the agent as a cause in bringing about some upshot by tracking the transference of this quantity, and assessing whether or not the transference was successful. For instance, when I throw a rock, I could say that my kinetic energy is transferred to the rock, and from the rock to the window, thus establishing a proper causal relation. On the other hand, my not fixing the boiler does not transfer any kind of quantity to the boiler: to be fair, one could argue that when I do not check my appliance the limestone accumulates in the pipes, the limestone in the pipes blocks the cooling system, the heat transfers to the back wall of the boiler and so on. This quantity, however, is not something the agent A passes to the object B, and the relation cannot be thus described as causal. Within this model, doing would then amount to causing, while allowing behaviours would not amount to causal connections.

Process theories seem to be immune from tricky counterexamples like preemptions. In Backup,

[8] Aronson's account is also taken to represent standard physical process frameworks in Schaffer (2000).

[9] These quantities are the ones used by Aronson as examples, but this is by no means intended as a closed list.

[10] Aronson (1971), p. 422.

any physical connection model delivers the classification of Alice's behaviour as doing, thus matching our natural intuitions in this case. Defining causation in terms of a physical connection between the agent and the upshot might also have some advantages in the light of the act/omission distinction. Process theories clearly do not have the problem of blurring what are intuitively acts and omissions, or doings and allowings. Doings are causes whereas allowings are not. But the problem remains of distinguishing omissions from causally unrelated events. According to Aronson's model, indeed, the boiler example would fall into the same category of the relation between brushing one's teeth and the water boiling, As I argued above, however, a causal model tracking our intuitions about human agency must recognise at least some omissions as causally relevant. Yet, the notion of physical connection fits well our intuitions of what counts as an "act", while, in simple counterfactual accounts, both actions and omissions can equally play the role of difference makers. A sophisticated process theory, which supplements this "core" or genuine definition of causation with further criteria accounting for omissions, might thus still have a better shot than counterfactual accounts. I now examine one of these refined proposals.

While many physical connection models bite the bullet and do not provide a detailed account of omission as causes (as Aronson himself, Armstrong (2004), or Beebee (2004)), other authors try to define a broader family of causal notions which can address those cases which are commonly perceived as allowings. Phil Dowe (2000, 2001), for instance, claims that even if omissions are not causes according to his conserved quantity theory, they nonetheless amount to a "close relative", which he variously defines as causation* or "quasi-causation", and whose identification relies on counterfactual analysis. Specifically, Dowe begins with an analysis of the crucial case of *prevention*:

> "A prevented B if A occurred and B did not, and there occurred an *x* such that: (1) there is a causal interaction between A and the process due to *x*, and (2) if A had not occurred, *x* would have caused B."[11]

This definition leads to the following account of omissions as "quasi causes":

> "not-A quasi-caused B if B occurred and A did not, and there occurred an *x* such that (1) *x* caused *B*, and (2) if A had occurred then A would have prevented B by interacting with *x*."[12]

In this sense my not checking the boiler (not-A), quasi caused the boiler to break (B), because the deterioration process of the pipes (*x*) caused B, and had I checked the boiler (A), this would

[11] Dowe (2001), p. 221.

[12] Ibidem, p. 222.

have prevented B. As results from the definition of prevention, it also holds that my checking the boiler would have prevented it from breaking: checking the boiler (A) prevents the break (B) because the associated counterfactual is true (if I check the boiler, it is less likely that it breaks); the process of deterioration *x* is physically connected to A, and if I do not check the boiler, the deterioration causes it to break.

Relying on Dowe's analysis,[13] we could identify "real" causes, within some specific physical connection account, as acts, and thus instances of "doing". On the other hand, relations between an agent and an upshot which satisfy the definition above would amount to "quasi-causation": these cases, arguably, capture our intuitive understanding of omissions and thus classify as "allowing". This framework, which can be considered a hybrid view between genuine physical connection accounts and difference-making approaches, seems to adequately satisfy the requirement made at the end of 1.2, i.e., to distinguish between different ways or degrees a conduct can be causally relevant to an upshot.

Even if more sophisticated process theories seem to allow for a satisfactory act/omission distinction, the project of capturing our intuitive understanding of causal relations in terms of physical connections faces some serious difficulties. While the cause *versus* quasi-cause (or similar) might appear a promising distinction, it does not seem to adequately map onto cases that we intuitively regard as doings *versus* allowings. This is not only the case of tricky or "borderline" examples, like withdrawing aid or removing barriers, which pose a serious challenge to any doing/allowing account. As Schaffer (2000) argues, it is more the case that physical connection amounts to just one way of causing. If we narrow down instances of genuine causation to relations where actual physical processes are involved, many behaviours we intuitively perceive as actions, and which cannot be possibly described in terms of physical processes, would thus fall in the quasi-causation, or allowing, category. Schaffer defines these examples as instances of causation by *disconnection*, and he claims that their pervasiveness threatens the soundness of process theories as a whole.

Suppose, for instance, that I am a powerful mobster and I call one of my affiliates, ordering the killing of the major. It would be difficult to explain my contribution to the major's death in terms of the transference of a quantity, or property, from me to the major. There are, however, even more worrisome examples: to show how causation by disconnection is ubiquitous,

[13] Dowe's account is of course more elaborate than the brief overview I offer here; specifically, his definition of quasi-causation is further refined so as to accommodate cases of preemption and overdetermination. In this discussion, I take Dowe's proposal as representative of a class of physical connection frameworks addressing the notion of quasi- causation in terms of counterfactual analysis, such as, for instance, Persson's (2002) "fake causation" account.

Schaffer famously analyses the case of an agent firing a gun and killing another man.[14] At every step of the causal route, Schaffer argues, we can see that causation works not by physical connection but instead by disconnection: for instance, the firing of the bullet through the victim's heart intuitively causes it to stop. But, as Schaffer observes, heart piercings cause death only by disconnection, as the brain is kept alive by an influx of oxygenated blood, and heart piercings cause death by disconnecting this influx, resulting into oxygen starvation. The mobster and the firing gun case are thus cases of quasi-causation. Of course, one might press that quasi-causation is still a relation of causal relevance. Yet, this dismissal is extremely counterintuitive and untenable: firing a gun, in fact, appears a straightforward case of genuine causation or “doing”, and one we would unanimously agree upon; a causal account which confines genuine causation and doing conducts to cases of physical connections seems therefore inadequate for tracking our understanding of the doing/allowing distinction.

In conclusion, elaborate versions of physical connection accounts, such as Dowe's proposal, seem at first to provide a straightforward solution to the issue of correctly distinguishing the causal contribution of acts and omissions, *versus* lack of causal relation altogether. Specifically, instances of physical connections between an agent and an outcome seem to match our intuitive understanding of acts, or “doings”. Nonetheless, the category of “quasi” causal relations, which should account for causally relevant omissions, does not successfully map our intuitive understanding of allowing. The model, indeed, does not account for causation by disconnection. Given the results of this analysis, I turn in the following section to sophisticated difference-making accounts, to see whether they are up to this task.

1.4 Distinguishing acts and omissions: the counterfactual approach

As I argue in 1.2, counterfactual analysis appears to be underspecified when it comes to discriminating between the different causal impact that acts and omissions can have on an outcome. In short, in the simple counterfactual model I examine above, both acts and omissions can count as difference makers, and thus causes; while it is certainly an advantage that some instances of omissions can be recognized as causally relevant, this becomes a problem if we are interested in what distinguishes “doing” from “allowing” conducts.

Moral philosophers have attempted to fill this apparent gap in the counterfactual model of causation by proposing ways to distinguish acts and omissions. The various proposals can be

[14] Schaffer (2000), pp. 286–292.

carved up in the more traditional terms of act/omission or action/inaction, but are also referred to as positive/negative relevance; in this study, I do not advocate for one taxonomy over the others, as my interest is rather to examine whether the suggested dichotomy tracks our doing/allowing classifications.[15]

Some negative/positive relevance accounts, like Donagan's (1977), do not seem to provide a substantially different analysis with respect to general counterfactual frameworks. According to Donagan, we can test for the different ways an agent could be relevant to an outcome making reference to what would have happened if the agent had "abstained from intervening in the course of nature",[16] that is, had not acted or made any movement at the given relevant moment. If the upshot would have occurred anyway, had the agent not interfered with the course of nature, the behaviour amounts to allowing; if the upshot would have not occurred, had the agent not interfered, the behaviour amounts to doing. In spite of being intuitively appealing, this account does not make any significant progress in distinguishing omissions from causally unrelated events. This counterexample by Frances Howard-Snyder (2002), for instance, shows that cases of collective responsibility are impossible to deal with appealing to this interpretation of positive/negative relevance:

> "Suppose an SS officer, Franz, tortures someone to death. But this is standard practice in the Gestapo. If Franz had stayed home with a sore throat, or if Franz had never existed, his pal Hans would have done the torturing, in the same way, at the same time Franz did. (...) then Franz is negatively relevant to the victim's death by torture. That is, Franz merely allowed the death to occur".[17]

Yet, this conclusion strikes us as counterintuitive, as Franz's action is clearly a doing. This, however, is not the only problem of Donagan's account: even if Franz's action were an actual instance of allowing harm, his behaviour does not count as a difference maker at all with respect to the outcome; this framework thus does not adequately account for omissions.

A much better worked out proposal is Jonathan Bennett's (1995), which is known as the "most of the things she could have done" account. For an agent to be positively relevant to an upshot, Bennett requires that most of the ways she could have behaved would not have lead to the upshot, while she is negatively relevant if this condition is not satisfied. In this sense, when I refrain from stopping the slipping rock, or I am too careless to fix my boiler, I am causally relevant to the rock's crushing a person or to the boiler's breaking, since I could have stopped

[15] I also assume that all the different dichotomies put forward in the literature refer to this intuitive distinction I call here doing/allowing.

[16] Donagan (1977), chapter 5.2, "Doing Evil".

[17] Howard-Snyder (2002), "The Doing/Allowing Distinction", Stanford Encyclopedia of Philosophy.

the rock or fixed the boiler; most of the ways I could have behaved, however, including going on holiday, having lunch, or brushing my teeth, would have lead to the same upshot. My behaviour is thus negatively relevant to the harmful consequence, and can be classified as an instance of allowing. To sum up, Bennett understands the positive/negative distinction in terms of how informative a proposition is about the agent's bodily movements. In the fixing my boiler example, the proposition describing my brushing my teeth, or going on holiday, is not particularly informative with respect to the outcome “the boiler breaks”. The behaviour thus counts as negatively relevant. On the other hand, throwing the rock amounts to a informative proposition with respect to the outcome “the window breaks”. Therefore, my behaviour counts as positively relevant to the outcome.

The first advantage of Bennett's framework is that it replaces the obscure notion of “course of nature”, absent the agent, with the slightly more tangible concept of the set of behaviours available to the agent. The specific relation among the set of all possible conducts, the behaviour displayed by the agent and the occurrence of the upshot is further analysed by Bennett in terms of *explanation*: one's behaviour is negatively relevant to an upshot if a negative fact about this behaviour is the least informative fact that suffices to complete a causal explanation of it; on the other hand, one's behaviour is positively relevant to that upshot if a positive fact about the latter is the least informative fact about one's conduct that suffices to complete a causal explanation of it. This intuition seems particularly adequate for many doing/allowing examples: if I don't check my appliances and the boiler breaks, the breaking of the boiler could be explained by the fact that I was brushing my teeth (a positive fact about my behaviour), but could also be explained by the fact that I did not check the boiler, that is, a negative fact about my behaviour. On the other hand, if I throw a rock at the window, a positive fact about my behaviour, that is, me throwing the rock, is sufficient for explaining the breaking of the window. But while Bennett seems to avoid reference to the normal course of nature, what counts as an “adequate explanation for an outcome” subtly turns on this notion. Suppose that a lifeguard on duty, instead of patrolling the beach, is partying with her friends. A child gets in the water, swims too far away, screams for help, nobody rescues her, and she eventually drowns. In this example, it looks like two “explanations” of the outcome are reasonably tenable: the child drowned because the lifeguard did not rescue her (negative fact, allowing), and the child drowned because the lifeguard was partying with her friends (positive fact, doing). In the next chapter, I argue that appeal to the normal or standard course of events is indeed inescapable when it comes to making doing/allowing attributions.

Beyond this observation, Bennett's more detailed (if ultimately incomplete) account of the doing/allowing distinction has the crucial advantage of distinguishing between omissions and causally unrelated events while, at the same time, allowing for omissions to be causally relevant. In Bennett's model, omissions can indeed amount to part of the explanation for the occurrence of a given outcome. Yet, they differ from acts insofar as they can be characterised as negative facts about the agent's behaviour.

Bennett's proposal is not itself immune to counterexamples, in which the suggested positive/negative relevance model seems to fail to keep track of our common understanding of behaviours as doings rather than allowings. Quinn, for instance, argues that in some cases Bennett's account wrongly classifies "not moving" as positive, while making any move whatsoever as negative. Therefore, his account fails to classify as doing harm some behaviours which intuitively fall in this category. To see this point, Quinn (1998, p. 295) discusses the following case:

> Reverse Immobility. Henry is in a room with a motion detector, which is connected to a bomb's detonator. If any motion is detected within the next minute, a bomb will go off in another room, killing Bill. If Henry remains perfectly still for one minute, the bomb will not go off, and Bill will live. Henry waves his arm and Bill is blown to smithereens.

In Reverse Immobility, it looks like Henry did make Bill die. However, on Bennett's account, Henry has merely allowed Bill to die, since the proposition "waving one's arm" is not particularly informative with respect to the outcome "Bill is blown to pieces". Arguably, Quinn claims, this is incorrect. I do not think that this counterexample is particularly compelling. Again, what strikes me as counterintuitive about this scenario is the peculiar situation the agent finds herself in, which makes the "standard" consequences of her behaviours very different from what could have been expected in "normal" cases.

To sum up, more elaborate counterfactual accounts, like Bennett's analysis, seem to have, despite some remaining difficulties, adequate resources for distinguish between the different causal impact of acts and omissions, and thus between doing and allowing behaviours. Nonetheless, we have so far identified a problem for counterfactual accounts of the doing/allowing distinction – there is a need to appeal to a further concept such as the natural course of events. In the remainder part of this chapter, I address what looks like a further challenge; I argue that responding to this challenge also ultimately rests on a notion of the normal course of events.

1.4.1 Not all omissions are allowings

Up to now, we have required from an adequate causal account to consider omissions as causally relevant, even if at a different level or degree with respect to actions. Specifically, I have suggested that it is both intuitive and practical to talk about omissions as causes, or at least as causally relevant. Equally sound motives, however, demand that not *all* omissions should count as causes. To illustrate this point, let's take the famous Neighbour/Queen of England example:[1]

> I'm leaving for a holiday break and I ask my neighbour to water my flowers; she forgets to do so and my flowers die.

Bennett's counterfactual account delivers in this case the intuitively correct classification of my neighbour's behaviour as an instance of allowing: most of the ways she could have behaved, indeed, excluding "water the flowers", would likely have lead to the death of the flowers anyway. The same intuition, clearly, does not hold for any other person's, such as, say, the Queen of England's, relation with my flowers. Counterfactual analysis, nonetheless, delivers exactly the same verdict about the Queen of England's behaviour which, just like my neighbour's, should thus count as "allowing" my flowers to die: after all, according to Bennett's interpretation, a negative fact about the Queen's conduct (not watering my flowers) also amounts to the least informative negative fact about her behaviour that explains the death of my flowers.

This counterexample is clearly worrisome: while it seems fair to accept that the neighbour's sloppiness had an impact on the death of the flowers, it appears ridiculous to claim that the Queen of England counts as a difference maker with respect to this upshot. This same reasoning, besides, can be extended, leading to the conclusion that I am causally relevant to the death of all flowers in the world (or at least in the neighbourhood) that I did not water. In short, we do not perceive all omissions as instances of "allowing" the upshot to occur. An adequate account of the act/omission distinction, therefore, must also properly discriminate between "genuine" omissions, which count as causally relevant, such as my neighbour's carelessness, and "false" omissions we do not perceive as causally relevant and thus proper instances of allowing, such as the Queen's minding her own business.

This issue, which is crucial for correctly tracking the doing/allowing distinction, amounts to a general challenge counterfactual analysis faces when it comes to justifying our intuitive judgements of causal relevance. In 1.2.1 I diagnosed this problem as one concerning context-dependency. A related issue that others address concerns the distinction between causes and *background conditions* (see, e.g., Schaffer (2005)). Considering a simple example that does not involve the doing/allowing distinction: when I drop a lighted cigarette which produces a fire, the

[1] Schaffer (2000), p. 297. This example is also discussed in detail in Hitchcock (2001).

lighted cigarette is taken to be the cause of the fire, while the presence of oxygen, which is equally necessary for the fire to develop, is considered as a fixed background condition. Similarly, the fact that the Queen of England does not usually care about my flowers could count as a fixed background condition of the context, while my neighbour is the relevant cause. This division, plausibly, explains why we consider the Queen's behaviour as not causally relevant, as opposed to my neighbour's negligence. As Schaffer observes, however, the selection we make between what is causally relevant and what is just part of the normal description of the context seems extremely arbitrary, and the task of spelling out adequate criteria has resisted many theoretical efforts. As Lewis (1986, p. 162) puts it:

> "We sometimes single out one among all the cause of some event and call it 'the' cause, as if there were no others. Or we single out a few as the 'causes', calling the rest mere 'causal factors' or 'causal conditions' (...) We may select the abnormal or extraordinary causes, or those under human control, or those we deem good or bad, or just those we want to talk about. I have nothing to say about these principles of invidious discrimination."

I focus here on Schaffer's proposal for solving this impasse. The proposal rests on the insight that we should look at causation not as a binary relation of the form "c causes e", but rather as a quaternary and *contrastive* relation of the form "c rather than c^* caused e rather than e^*".[2] Specifically, Schaffer proposes the following counterfactual definition for contrastive causation: "c rather than c^* causes e rather than e^* if and only if, if c^* had occurred, then e^* would have occurred (formally, $O(c^*) > O(e^*)$) , where C^* and E^* are non-empty sets of contrast events". That is, C^* and E^* are sets of variables that may be represented by a set of values. The second main insight behind the idea of contrastive causation is that the sets C^* and E^* are fixed by *context*. So, when asking "what initiated the fire?", Schaffer suggests that we set the presence of oxygen as a fixed background condition and focus on, for instance,

{c: the dropping of a lighted cigarette; c_1^*: the occurrence of a shortcut; c_2^*: a lightning strike},

as only c_1^* and c_2^* amount to relevant alternatives. In the same way, when asking "why did the flowers die?", the event that the Queen could have watered my flowers amounts to such a remote possibility that it should not be included in the set C^* of the relevant alternatives with respect to the context. Arguably, different sets C^* will be triggered, depending on the context in which the causal inquiry is made. Yet, as Schaffer admits, a proper solution to the issue of causal selection should further elaborate adequate criteria for distinguishing between relevant alternatives and fixed background conditions.

The selection of real omissions from false omissions, in this sense, seems to share the same kind

[2] Schaffer (2005), p. 297.

of difficulty discussed at the beginning of this section in relation to the positive/negative relevance distinction proposed by Donagan and Bennett. An adequate causal account of the doing/allowing distinction, as argued so far, requires that a) we can properly distinguish acts from omissions, and b) we can properly distinguish real (or causally relevant) omissions from false (or not causally relevant) omissions. In a difference-making analysis, the complete success of both tasks seems to rely, however, on notions such as the natural course of events, the normal or salient explanation, the identification of a standard set of background conditions and the like. On the one hand, by allowing an agent's impact on an outcome to come in different degrees (the extent to which the probability of the effect is raised), counterfactual analysis appears thus to accommodate the doing/allowing distinction. On the other hand, it appears that a supplementary account of what is the standard or normal course of events is needed, if the counterfactual approach is to deliver the correct verdicts regarding the doing/allowing distinction. In other words, counterfactual reasoning, as exemplified by elaborate accounts such as Bennett's or Schaffer's, has the necessary tools to identify actions and omissions in a way that matches our doing/allowing classifications, thus satisfying both a) and b). In order to get the right results, however, we need to better elaborate the criteria for assessing which conduct is normal more appropriate for the agent in a given situation.

In the first half of Chapter 2, I will discuss the main competitor to causal accounts of the doing/allowing distinction, which I call "norm-based accounts". In the latter part of Chapter 2, I will show how insights from these so-called norm-based accounts can in fact help to fill in this remaining gap in the counterfactual approach. In this respect, I will argue that these accounts are better seen as supplements to the counterfactual approach.

Before examining the norm-based accounts, and how they can contribute, rather than compete with, difference-making accounts, it is worth making sure that a sophisticated difference-making account, supplemented with an appropriate interpretation of what counts as a "normal" behaviour or standard interpretation, is really the most suitable alternative on the table for tracking the doing/allowing distinction. To this end, I want to address one last class of theories that have been suggested by moral philosophers. I will call these "informal process theories" or "sequence accounts".

1.5 "Sequence" Theories

In discussing process theories, I argued that physical connection does not always track our ordinary judgments about which behaviours amount to positive actions, or genuine instances of causation. Some moral philosophers interested in the doing/allowing distinction, however, suggest what looks to be a "looser" interpretation of physical connections so as to accommodate disconnection counterexamples, and match our intuitive act/omission distinction. Note that, if these accounts are indeed successful, they would amount to a better alternative than elaborate difference-making proposals, as they would account for the doing/allowing distinction without appealing to the notions of normal behaviour or standard explanation.

In moral theory, less stringent interpretations of process theories, which do not require the actual transference of properties or energy, have been put forward to better cope with the notion of human agency. Foot (1978, 1984, 1985), for instance, argues that the agent's role in bringing about a harmful effect is to be classified as initiating, sustaining, enabling or forbearing-to-prevent an appropriate harmful *sequence* which directly connects the agent to the upshot. These different levels of contribution to the upshot determine the classification of a conduct in terms of doing rather than allowing. Philip Wolff (2007, pp. 85–89), on the other hand, suggests an account based on the "configuration of forces". Within this model, causal relations between an *affector* and a *patient* are dependent on the following three conditions: 1) whether the patient is represented as having a force-based tendency towards an endstate, 2) whether the forces represented as exerted by the patient and affector are concordant, and 3) whether the patient is represented as making progress towards the endstate. For instance, according to Wolff, the notion of "causation" is applied if a patient does not have a tendency towards an endstate, the affector exerts a force on the patient towards the endstate (the forces of the patient and affector are not concordant), and the patient makes progress towards the endstate. "Doing" actions, Wolff argues, satisfy all the conditions above: when the assassin fires the bullet, the victim does not have a tendency towards the endstate "dying", the bullet fired by the gun exerts a not concordant force with respect to the oxygen flux, and the patient thus makes progresses towards the endstate "dying". On the other hand, "allowing" cases occur when the patient has a tendency towards the endstate, the forces exerted by the affector and the patient are concordant and the patient progress towards the end states. These conditions, for instance, apply to the classic "allowing" case of the slipping rock. Barry and Øverland (2017), similarly, identify two criteria so as to assess whether the conduct amounts to doing or allowing: whether the behaviour of the agent is a relevant action (i.e., the agent has a direct enterprise in bringing about the upshot) and whether there is a complete causal process connecting the agent and the upshot.[3]

Fiona Woollard (2015) suggests a more sophisticated model, which draws on both Foot's and

[3] Barry and Øverland argue that, according to this model, the third category of "enabling" should supplement the classic doing/allowing dichotomy.

Bennett's proposals. As Woollard's account ultimately appeals to the notion of harmful sequence, I discuss it in this class of refined process theories.[4] According to Woollard, Foot's analysis must be supplemented with the idea that we should look at whether a relevant fact about the agent's behaviour is part of the harmful sequence. Woollard first argues that "an agent counts as doing harm if and only if some fact about the agent's behaviour is part of the sequence leading to harm; the agent counts as merely allowing harm if and only if a fact about the agent's behaviour is relevant to, but not part of, this harmful sequence."[5] However, this definition must be supplemented with an account of what makes a fact a part of the harmful sequence. To do so, Woollard distinguishes between *substantial* and *non-substantial* facts: the former are more intuitively perceived as "natural" parts of a sequence, such as my throwing a rock, while the latter usually amount to fixed background conditions, such as my brushing my teeth every morning. Specifically, a fact counts as substantial if it is either positive or contradicts normal presuppositions.[6] She then concludes that if an agent is relevant to a harm through a non-substantial fact about her body, then her actions are merely a condition for, rather than part of, the harmful sequence, and this would count as allowing harm. On the other hand, if there is a complete sequence of substantial facts leading from the agent to an harmful effect, the agent's action would count as doing. This account seems to fit our intuitions about what counts as doing and what counts as allowing. There is a harmful sequence in place connecting my neighbour to the death of my flowers. My neighbour's action also appears to be a condition for the harmful outcome: she is relevant to the upshot through a negative (and thus non-substantial) fact about her body, that is, not watering my flowers, and her behaviour thus counts as allowing. On the other hand, a man firing his gun counts as part of the harmful sequence: there is a positive fact about his body (firing the gun) which connects him through a complete chain to the harmful upshot.

All these accounts, in short, stick to the intuition that there must be a sort of physical connection or succession of events relating the agent and the outcome, but further examine the agent's behaviour on the basis of the *type* of interaction between the agent and this causal sequence; the doing/allowing distinction tracks the different nature of these interactions, which stand for different degrees of the agent's causal contribution with respect to the outcome.

A clear advantage of these models is that they replace a very narrow account of what counts as a

[4] Note that Woollard does not define "sequences" in causal terms. Specifically, she argues that "sequences are not spatio-temporally continuous causal chains." (p. 28). It is thus in a broad sense only that I discuss Woollard's account among informal process theories, insofar as she uses the notion of a chain of facts connecting the agent to the outcome.

[5] Woollard (2015), p. 23.

[6] Woollard's definition of positive/negative facts draws on Bennett's analysis. For a more detailed survey of the different ways in which a fact can be substantial, see Woollard (2015), pp. 36–59.

“connection” with a broader, and more intuitive, notion of a natural sequence or succession of events. As Wolff's interpretation of the bullet case suggests, there is a very prominent and tangible connection between the fired bullet and the death of the victim, besides the hard facts about the physical processes going on at the level of human biology. Similarly, it seems extremely reasonable to describe my flowers dying as a natural and self-sustaining sequence of events that my neighbour appears to “allow”. The notion of “relevant fact”, likewise, appears to conveniently track our intuitions of “doing something in a causal way”.[7] Arguably, these proposals can successfully cover most of our attributions of doing and allowing, by relying only on this apparently self-evident and macroscopic identification of succession of phenomena, and on the agent's contribution to the sequence.

My worry, however, is that these models are just concealing and making implicit the appeal to what counts as normal or standard, especially when it comes to sequences and configurations of forces involving human agency. Specifically, I argue that some prior judgements over which explanations are salient, which behaviours are appropriate, or what amounts to a positive action lie behind the identification of a certain sequence, configuration of forces among events, or even “tendency towards an end-state”[8]. Moreover, while processes such as a rock slipping might stand out as “natural”, or sequences such as drowning a person seem highly discernible, picking out the appropriate sequence is not as straightforward when evaluating, for example, my neighbour's contribution to the death of the flowers. Arguably, the reason why my neighbour's role with respect to the upshot, as opposed to the Queen's, is conceived of as “allowing”, is that there is a general understanding that, once a promise has been made, this generates certain obligations in the promiser with respect to the promisee. This, in turn, identifies the neighbour as contributing to the harmful sequence, in the sense of failing to interrupt it, but not the Queen, who is perceived as completely extraneous to this sequence. In short, just as we can pick out a sequence connecting my neighbour to the flowers, we could also look at the sequence connecting the Queen to the flowers; clearly, only one sequence seems appropriate when assessing causal relevance, but this appears to be a matter of choice rather than of hard facts. Woollard seems to make a similar remark when discussing the distinction between substantial and non-substantial facts, and ultimately concedes that these classifications, which make a substantive work in distinguishing doings and allowings, rely on prior judgements.[9] My point is that, in selecting the adequate sequence and assessing the agent's contribution to it, these accounts already encode a specific understanding of which explanation is relevant for the outcome to occur, and which individuals could or could not intervene on the sequence. This,

[7] Barry and Øverland (2017), pp. 82–83.

[8] Wolff (2007), p. 87.

[9] Woollard (2015), p. 61.

however, reflects a previous evaluation of which behaviours are expected, which explanations are commonly accepted and which kind of obligations and social interactions are at play in the context at issue.

In this sense, Foot-style accounts of the doing/allowing distinction, while extremely plausible and intuitive, still depend on this prior assessment of what amounts to a normal sequence of events, just like elaborate difference-making accounts like Bennett's or Schaffer's. With respect to the latter, however, these models seem to have the disadvantage of making this feature less explicit and, arguably, less debatable and identifiable. I conclude, therefore, that a framework that incorporates and makes explicit the prominent and central role of norm-based considerations is preferable for reasons of clarity and transparency.[10]

[10] A notable exception is Woollard's (2015) proposal, which discusses at length the substantial/non-substantial distinction. As I will discuss in Chapter 2, however, we can assign to this feature an even more prominent role in cashing out the doing/allowing dichotomy.

2. Analysis of Doing *vs* Allowing: "Norm-based" and "Mixed" Accounts

In this chapter, I continue my survey of the most adequate accounts for tracking our every-day use of the doing/allowing distinction. I introduce the main alternative to causal accounts; I call this the "norm-based" account. I then argue that a suitable model for the doing/allowing distinction should incorporate aspects from both these two frameworks. Specifically, I describe in detail Christopher Hitchcock's "self-contained network" account, and illustrate its merits in matching our intuitive doing/allowing attributions, whilst at the same time solving most of the difficulties analysed throughout Chapters 1 and 2.

2.1 Norm-based accounts

In this chapter I use the label "norm-based" to describe all accounts that attempt to justify the doing/allowing distinction by referring as well to independent and external features and principles, which are usually of moral nature. Note that these accounts do of course retain some causal and/or descriptive aspect; after all, they rely on a notion of when an agent acts. But the difference with causal accounts is that when it comes to classifying different actions as doings rather than allowings, there is a more direct direct appeal to norms rather than an attempt to appeal to causal notions. For the purpose of this chapter, I only focus on whether these models adequately match our common doing/allowing classifications, and can thus make sense of our intuitive use of this distinction. In this section, my aim is not just to provide a survey of these accounts, but also to see whether some aspects of these positions may be used to flesh out a more elaborate difference-making model, in the sense of supplementing it with a more detailed understanding of what is "normal".

I first deal with the most influential of these accounts, which I call the rights-based proposal, analysing Philippa Foot's (1978) and Warren Quinn's (1989) proposals; secondly, I turn to Shelly Kagan's (1989) framework, which can be considered as a broader norm-based proposal. Note that, in this preliminary survey, I am concerned specifically with how different accounts of the doing/allowing distinction draw the line between doing and allowing behaviours, and not with the further issue of justifying whether this distinction matters morally. For this reason, I define here as "norm-based" accounts only the frameworks which appeal to notions of self-ownership, rights, and rules of conduct as a way to discriminate between doing and allowing behaviours, and not the proposals which use these notions to further explain moral significance.

2.2 Rights-based accounts

I interpret Philippa Foot's (1978) account as the first “rights-based” analysis of the doing/allowing distinction. She argues that negative rights (namely, rights against one's interference) are different from, and stronger, than positive rights (rights to be provided a good, or aid). She further argues that we classify an action as “doing” when it violates one's negative rights, while we classify an action as “allowing” when it violates a positive right. Note that, in Chapter 1, I classified Foot's account as an informal process theory, insofar as it rests on the idea of a sequence connecting the agent to the upshot. Nonetheless, Foot draws a line among the different ways an agent can be relevant to a harmful sequence, and classifies initiating and sustaining a sequence as doing, while enabling and forbearing-to-prevent amount a sequence as allowing. Therefore, while the distinction Foot draws among four different ways of being related to a harmful outcome is purely descriptive, the further distinction between which sequence counts as doing and which as allowing is not merely descriptive. This further distinction, in this sense, is based on the insight that initiating and sustaining involve a violation of negative rights, while enabling and forbearing-to-prevent involve a violation of positive rights. Foot's account, therefore, retains a causal aspect, but then makes use of the concept of negative/positive rights as to draw the doing/allowing distinction.

Warren Quinn's proposal provides a similar analysis of the relation between doing/allowing and positive/negative rights. Quinn's reflection begins with the insight that the most simple and straightforward way to think about the doing/allowing distinction, the action/inaction distinction, is open to counterexamples. In general, reviewing Bennett's and Foot's accounts, he notices that no matter how sophisticated or refined a causal analysis of the doing/allowing distinction is, there seem to be some cases we perceive as doing, but which nonetheless fall into the “allowing” category according to the model at issue. In particular, Quinn claims that in some instances we undoubtedly kill, or harm, without having caused the harm to occur, having positive relevance to the consequence and so on.

Suppose, for example, that I am driving a train directed towards a house on fire where people are trapped, because I want to rescue them. Ahead of me, I notice someone tied to the tracks, and I decide not to stop the train and so run over the person.[1] This action, Quinn argues, amounts to killing (or doing), despite my good intentions to save the people trapped in the fire and, most of all, despite the fact that most causal models would not describe my behaviour as causing the death, or as being positively relevant to the death. For instance, on Bennett's terms, the agent could count as only negatively relevant to the death of the victim, since most of the way she could have behaved would have resulted in the train running over the person. Specifically, it looks like a negative fact about the agent's behaviour − not stopping the train − is part of the explanation of the harmful upshot; there is seemingly only one way the agent could have behaved, that is stopping the train, which would have prevented the death of the person.

The reason why we regard this action as an instance of doing, therefore, must lie elsewhere, and amounts to something more than what can be captured by a factual analysis of the events involved. Quinn's intuition is that the person tied to the tracks has authority or ownership over her body, which creates a specific obligation in other agents towards her, and a general expectation that this authority is to be respected. In other words, Quinn is speaking the language of *rights*: some rights a person has, such as “I have the right not to be drowned”, are *negative* rights, which generate an obligation not to interfere with the person, and cannot thus be trumped. On the other hand, some rights are *positive*, in the sense that they require the intervention of some other agent for their fulfilment, such as “I have the right to be rescued from drowning”. Positive rights certainly generate some sort of obligation, but, Quinn argues, they are clearly less compelling and stringent than the kind of obligations negative rights give rise to. The subsequent move is then to equate cases of “doing” with violations of negative rights and cases of “allowing” with violations of positive rights. This account, therefore, is similar to Foot's, but it extends it with an underlying justificatory concept of self-ownership, from which the notion of rights stems. To be clear, Quinn also offers a descriptive account of the doing/allowing distinction; my argument is rather that most of the focus and of the theoretical work is done at the level of the concept of self-ownership.

Quinn's proposal, as Woollard (2015, p. 107) argues, seems extremely persuasive: ownership over one's body and mind could thus be an intuitively relevant feature when judging an agent's conduct as an instance of doing harm rather than allowing it to occur. Specifically, Woollard argues that what we are concerned with when making this distinction is the fact that we are *imposing* on others: an agent is doing harm if she intrudes upon what belongs to others. In the train example, the person tied to the track has ownership over the integrity of his/her body, such that the fact that another agent runs over me makes the action an intrusion or imposition.[2]

[1] For a detailed discussion of this example, see Quinn (1989).

[2] To be clear, I do not consider Woollard's proposal as a norm-based account, insofar as it only uses the notions of self-ownership and imposition to justify the moral relevance of the doing/allowing

Rights-based justifications of the doing/allowing distinction have been variously criticised with the charge of *arbitrariness*. Quinn's account, indeed, like Foot's, selects a specific moral feature, such as the distinction between negative and positive rights violation, or the related concept of self-ownership, and assumes that this distinction determines the classification of an action as doing or allowing. But, as Howard-Snyder (2002) objects, there is nothing to the positive/negative dichotomy that makes it "trump", or makes it particularly significant in assessing moral value, or, more accurately, in distinguishing between doing and allowing. Why would other classifications, such as "rights of adults" and "rights of children" be not equally suitable candidates for informing our moral judgements? The fact that we rely on a specific moral norm rather than another in spelling out doing/allowing, in short, seems extremely arbitrary and unjustified, because it requires us to commit to a specific moral theory. Moreover, the problem of providing an adequate justification for the doing/allowing distinction is then simply shifted to the next level, namely to the problem of why the positive/negative rights distinction, among others, is particularly relevant. This consideration leads to an even more serious problem for rights-based accounts, which is their lack of explanatory power in justifying the significance of the doing/allowing distinction. By equating doing and allowing with negative and positive rights violations, indeed, we are simply re-labelling an already existing morally relevant dichotomy: in short, we already know that violating rights is morally bad, and violating negative rights is comparably worse than violating positive rights. In this sense, these doing/allowing models are simply introducing a new pair of terms, which completely overlap with positive/negative rights violation. But then there is nothing distinctly morally relevant about doing *versus* allowing.

This line of criticism, while serious and challenging, is nonetheless not the focus of this chapter, which is instead the assessment of alternative analyses of the doing/allowing distinction. The charge of arbitrariness, indeed, raises the further issue of whether any norm-based account could in principle provide an adequate justification for the distinct moral significance attached to doing and allowing. Here, however, my task is just to test whether a specific account can consistently track our common doing/allowing classifications.

In this respect, the standard theory of positive and negative rights violation, which Woollard spells out as "not intruding", seems to match our intuitions in cases which are, at least to some extent, "ethically sensitive": the trolley-style example used by Quinn, or the Drowning/Rescuing cases. Seemingly, however, for some "non moralised" examples, the discussion in terms of the violation of rights may not turn out to be so reasonable. Let's take, for

distinction. In this sense, I only refer here to her discussion of the merits of Quinn's account, and to how she elaborates on Quinn's concept of self-ownership.

instance, the example of the Neighbour: while there could be agreement over the fact that she is allowing my flowers to die, it would be difficult to justify this classification in terms of rights violation; it would, indeed, seem excessive to claim that I have a positive right to have my flowers watered during my trip, or that such an informal promise is kept. Similarly, cases where the (harmful) upshot amounts to a minimal threat or even nuisance to other people may be difficult to describe as intruding or imposing over these people's belongings or safety. Nonetheless, the kind of behaviour performed by the agent could strike us as an instance of "doing" or "allowing" harm.

Another counterargument to the rights-based account is Shelly Kagan's (1989, p. 101) respirator example: in case a), my enemy sneaks into the hospital room where I lie in a coma and removes my respirator; in case b), by contrast, the doctor, acting upon the decision of the ethical committee, switches off the respirator. Kagan argues that case a) amounts to doing harm, while b) to allowing a harm to occur. This classification, which is seemingly reasonable, cannot thus be easily captured in terms of negative or positive rights, or in terms of "not intrusion": the distinction, rather, seems to appeal to further ethical norms, principles and considerations.

In summary, the rights-based account claims that the doing/allowing distinction tracks positive/negative rights violation, or whether there was an imposition on something in the realm of the agent's self-ownership. While these accounts seem generally reliable in tracking our doing/allowing assessments, there are some cases that are not well captured by the negative/positive rights dichotomy. First, in these cases, it could be argued that other expectations or obligations, which are less stringent and binding than rights or intrusion, may guide our doing/allowing assessments, like in the Neighbour example. This would not preclude the rights-violation account from serving as a specific or narrow version of a broader model. Secondly, there are cases where an agent's behaviour violates someone's positive rights, but in a way which intuitively characterises as an action. If I stop the lifeguard who is coming to rescue you, I am only interfering with a positive right of yours, that is, the right to be rescued; nonetheless, it is reasonable to argue that I "did" harm. Plausibly, it still appears to be relevant, in distinguishing between doing and allowing, the way an agent behaved in relation to the outcome, in terms of impact and causal relevance. In this sense, it may not be reasonable to completely dismiss causal models as a tool for tracking the doing/allowing distinction; rights-based accounts, thus, may be more fruitfully employed as a complement rather than an alternative to causal accounts. Before turning to this hypothesis, I will consider whether there is a better norm-based account which can accommodate for these difficulties.

2.3 Kagan's norm-violation account

In his survey of the debate concerning doing/allowing, Shelly Kagan (1989) notices that it is extremely difficult to find a univocal explanation that consistently and reliably captures all our classification of actions as doing or allowing harm, especially when it comes to "gimmicky cases" and fictional counterexamples. His idea is that this difficulty can be traced back to the fact that, when we evaluate whether an action is an instance of doing or allowing, we make reference to a set of norms: when an action violates one of these norms, this becomes a *salient* feature in the process of bringing about harm, and it is thus characterised as doing harm. If, on the other hand, the behaviour that brings about harm complies with the set of commonly accepted norms, we perceive it as allowing harm. With respect to the rights-based account, Kagan suggests a different analysis of the doing/allowing distinction: within the former view, both doing and allowing harm amount to rights violations, but doing actions would capture violations of "more important" rights. On the other hand, Kagan's proposal is that only doing actions involve a kind of norm violation. As to the concept of norms, Kagan allows for two alternative interpretations, one merely descriptive and one normative. In the former interpretation, norms simply refer to what are "normal" or common behaviours in specific circumstances. On the other hand, we can regard the concept of norm as normatively loaded, thus referring to moral principles and rules we ought to conform with.

This account appears to offer a convincing psychological justification for our common use of the doing/allowing distinction: something may strike us as an instance of doing something when it is "abnormal", and contrasts with widely accepted rules of behaviour. This idea, which is underlying Bennett's and Schaffer's models, also plays a central role in Hitchcock's account, discussed in the following section. Note too that this framework does not force us to commit to any specific moral theory or principle, thus avoiding the charge of arbitrariness. While rights-based accounts make substantial claims about the content of moral norms, this framework does not rely on any specific content or structure of such norms, but merely on the basic concept of norm compliance *versus* violation.

We can now turn to assessing whether Kagan's norm-violation account complies with our intuitions regarding the classification of cases. In many circumstances, this model seems successful in singling out commonly perceived "doing harm" actions as instances of norm-violation. Arguably, all the doing harm examples discussed throughout Chapter 1 can be cashed out as violations of some norms, or at least violations of standardly expected behaviours, such as breaking the window by throwing a rock. My worry, however, is that this framework is not as accurate when it comes to cases of allowing harm.[3] More precisely, Kagan's proposal suggests that, for any behaviour which results in a harm, if the behaviour violated a norm, then it is an instance of doing; on the other hand, if the behaviour complies with standard rules of conduct, it is an instance of allowing. This understanding, nonetheless, does not seem accurate without further criteria for identifying which norms or rules can carry out the test. To be sure, for some allowings behaviours, Kagan's account seems to deliver the right classification. If I do not save a person from drowning, and there are a lot of people around who could save her, it is reasonable to argue that I was not the one who was expected to do the saving, and thus my action characterises as allowing harm. But what about the Boiler or the Neighbour example? Standard rules of good conduct require that I regularly check and monitor my house appliances to prevent incidents and subsequent threats, or that my neighbour should keep a promise she made to me. In this sense, my negligence, or the neighbour's, should count as doing, a conclusion that is apparently counterintuitive. Of course, it could be argued that these norms are not as stringent as the obligation not to throw rocks, and this distinction explains the different assessments of the two cases. Kagan, however, does not offer any further details beyond discriminating doing *versus* allowing in terms of norm-violation.

My intuition is that the agent's causal impact on the upshot, and whether her behaviour was, so to speak, positively or negatively relevant, still has some influence over doing/allowing classifications. Imagine, for instance, that Bob is stung by a bee and is in anaphylactic shock. A doctor standing nearby has an adrenaline shot which could save Bob, but refrains from administering the injection. The doctor is certainly violating a fairly strict norm of behaviour, or her so-called “physician's oath”. However, it seems untenable to describe her conduct as doing harm to Bob: the sting caused the anaphylactic shock, and the doctor is just refraining from performing a medical procedure, and thus allowing the harm to occur, no matter how hideous her behaviour is. This judgment, arguably, captures the insight that the bee and the doctor have a different causal impact on the harmful upshot; it seems, therefore, that the doing/allowing distinction is also sensitive to this feature.

In conclusion, Kagan's suggestion, though not completely successful, is extremely insightful in relating the doing/allowing distinction to a salient abnormal or inappropriate conduct *versus* an ordinary but morally wanting conduct. While I agree that norm-violation has a prominent role in tracking doing/allowing classifications, different types of causal impact seem to play as well an irreducible part as well, which Kagan's proposal fails to account for.

In the next section, I suggest that the most adequate account for tracking the doing/allowing

[3] This is a worry Kagan is aware of too: he claims that this “norm-violation” interpretation might be too crude (p. 97), but he does not offer any indication so as to refine this account.

distinction relies on a difference-making model to assess the impact or the role of the agent, but also relies, in assessing the action/omission distinction, on the idea of which conduct is "normal" and appropriate for the agent, and thus on an underlying "norm-based" framework.

2.4 Hitchcock's "self-contained network" model

Most recent approaches in the literature on causal relations have employed structural equation frameworks to make sense of counterfactuals (Hitchcock 2001, 2007; Woodward 2003; Woodward and Hitchcock 2003, Halpern and Hitchcock 2013). In his 2007 paper, specifically, Hitchcock tackles, amongst other things, the issue of adequately discriminating acts from omissions, and argues that the idea of "self-contained networks" can successfully capture this distinction. While this model shares most of the features of counterfactual accounts, it can also be considered as incorporating many aspects of so-called norm-based accounts, especially in the assessment of the distinction between *default* and *deviant* variables.

Before discussing this central insight, I briefly sketch Hitchcock's structural equation framework.[4] First, let a *causal model* be an ordered pair <V, E>, where V is a set of variables and E is a set of equations among these variables. For simplicity, a variable here can take two values, where one value represents the occurrence, and the other the non-occurrence of a given event, or of a specific version of the event. Let's take this straightforward Assassination example: Alice poison's the victim's drink, and the victim dies. The variables in the story are:

$A = 0$ if Alice does not poison the drink, 1 if she does;

$C = 0$ if the victim does not die, 1 if she does.

Hitchcock argues that the counterfactuals we use when discussing the case (in Assassination, "if Alice had not poisoned the drink, the victim wouldn't have died") can be represented by equations among the variables: the variables on the the right-hand side of an equation, specifically, work as antecedents of the corresponding counterfactuals, while those on the left work as consequents. In Assassination, the equation describing the causal model is:

$C = A$

At this point, we can calculate the value of a variable on the left-side of the equation depending on the values taken by the variables on the right-side. For instance, for $A = 1$ that is, when Alice

[4] Throughout this section, I follow Hitchcock's (2007) formalisation.

poisons the drink, we have C = 1, that is, the victim dies. For the equation C = A, we can stipulate that C *counterfactually depends* on A, because we can compute the value of C fixing the value of A, and the resulting counterfactuals are true: if Alice had not poisoned the drink, the victim would have died; if Alice had not poisoned the drink, the victim wouldn't have died.[5]

For reasons of convenience, Hitchcock suggests that we can represent causal models as graphs, with nodes corresponding to the variables; an arrow from one variable to another represents the fact that the former appears on the right-hand side of an equation with the latter on the left. Hitchcock then defines the former variable as a *parent* of the latter. For Assassination, we thus have:

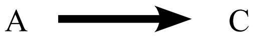


Figure 2.1: Assassination

Where A is a parent of C.

Let's now see how this simple model can be further refined so as to distinguish between acts and omissions, which amounts to the first *desideratum* spelled out in Chapter 1. To do so, let's take this second assassination example, which I call Bodyguard:[6] Alice poisons the victim's drink; the victim's bodyguard has an antidote but she does not administer it to the victim. Obviously, the victim wouldn't have died if Alice hadn't poisoned the drink, but she also wouldn't have died had the bodyguard administered the antidote. The causal graph representing this story is the following:

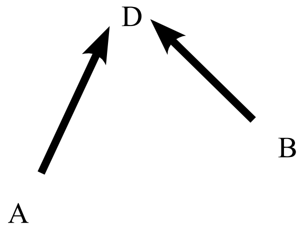


Figure 2.2: Bodyguard

Where:

A = 1 if Alice poisons the victim's drink, 0 if otherwise;

B = 1 if Bodyguard administers the antidote, 0 if otherwise;

[5] More technically, Hitchcock (2007, p. 502) defines the notion of counterfactual dependence as follows:
Let <V, E> be a causal model, let X, Y $\in$ V, and let the actual values of X and Y in the model be x and y, respectively. Y counterfactually depends upon X in <V, E> just in case there exist values of X and Y x' ≠ x, y' ≠ y (respectively) such that "if X had x', then Y would have been y'" is true in <V, E>.

[6] Ibidem, p. 504.

D = 1 if victim dies, 0 otherwise: and

D = A & not-B.

The difficulty with this case is the one of correctly identifying the causal impact of actions and omissions. Counterfactually speaking, indeed, Alice's poisoning the drink is *causing* the death of the victim in exactly the same way the bodyguard's refusing to administer the antidote is: both A and B are thus parents of D. This conclusion, of course, strikes us as intuitively wrong, suggesting that we need further criteria for distinguishing actions from omissions (which is the first *desideratum*). Recall that, as second *desideratum*, an adequate model of the doing/allowing distinction also has to distinguish between “real” and “false” omission.

With respect to the first task, Hitchcock argues that his model can successfully account for the difference between Alice's and the bodyguard's behaviours, by defining two alternative mechanisms causation can amount to, each capturing the specific way Alice and the bodyguard are causing the outcome. According to Hitchcock, in Bodyguard, when we read the counterfactual “had Alice not poisoned the drink, the victim wouldn't have died”, this appears to be a self-contained story, and Alice's behaviour seems a satisfactory explanation for the victim's death. On the other hand, when we read the counterfactual “had the bodyguard administered the antidote, the victim wouldn't have died” the story is not self-contained or complete: we feel we should know more, as refraining from giving the antidote would not itself and alone bring about the victim's death.

The idea of self-contained or else incomplete causal relationships, relies, in Hitchcock's view, on another distinction, the one between *deviant* and *default* values of a variable. The default value of a variable is defined as the value that the variable would take if there was no further information about intervening causes, and the situation were a sort of “self- persisting” system. For instance, in both Assassination and Bodyguard, the default value for C and D is 0, as it is reasonable to expect that, without anyone trying to poison her, the victim would stay alive. A variable which takes a deviant value, on the other hand, amounts to an event that somehow requires an explanation, like the fact that Alice decides to poison the victim's drink. Hitchcock claims that, in the realm of human behaviour, this distinction allows us to identify self-contained *versus* non-self-contained networks and thus track the act/omission distinction. As should be clear from this explanation, what default values we assign to variables depends on our experience and our judgment; it is not something that we can settle independently of our broader understanding of the situation.

Let's now see in more detail how deviant and default values can help in distinguishing between self-contained and non-self-contained causal networks. The idea is that we can think of self-contained causal networks as networks providing a "sufficient" explanation of the causal relation at issue. The connection between the drink being poisoned and the victim's death, in this sense, amounts to a satisfactory self-sustaining explanation of the events. On the other hand, a causal network is non-self-contained if it strikes us as incomplete: in short, to explain the occurrence of the outcome, we must appeal to other features which are not included in the network. For instance, the fact that the bodyguard did not administer the antidote is by no means a satisfactory explanation for the death of the victim. According to Hitchcock, we can think that a causal network is self-contained, when, if all the parents of a variable X all take their default value, they cannot cause X to take its deviant value. More intuitively, a causal network "is self-contained when it is never necessary to leave or augment the network to explain why the variables within the network take the values that they do. When a variable (...) in a self-contained network takes a deviant value, this can be explained in terms of the deviant value of one or more of its parents in the network."[7]

Let's be more precise here about what counts, according to Hitchcock, as a causal network. First, Hitchcock introduces the notion of a *path* as the "set of variables that are all connected by a series of arrows that meet tip to tail." In short, we can think of paths as a causal "route" which leads from a parent(s) to a child. In Assassination, there is only one path connecting A and C, namely {A, C}. In Bodyguard, {A, D} and {B, D} are the two causal paths connecting A with D and B with D respectively, that is, two possible "routes" or ways an outcome can be produced. A causal network connecting variable X with variable Y can then be defined as the set of all variables that feature in paths connecting X to Y. In both these simple examples, the causal networks coincide with the paths: the causal network connecting A with C is {A, C}, while {A, D} and {B, D} are the causal networks connecting A with D and B with D.[8]

We can now define more formally when a causal network is self-contained *versus* non-self-contained. Hitchcock provides the following definition, which captures the idea of "sufficient" explanation expressed above:

> "Let <V, E> be causal model, and let X, Y $\in$ V. Let N $\subseteq$ V be the causal network connecting X to Y in <V, E>. Then the causal network N is self-contained if and only if for all Z in N, if Z has parents in N, then Z takes a default value when all of its parents in N do (and its parents in V\N take their actual values)."[9]

By implication, a causal network is non-self-contained if and only if, for some Z in N where Z

[7] Hitchcock (2007, p. 510).

[8] I distinguish here between these two notions as, to discuss the case of preemption, I will appeal to both.

[9] Hitchcock (2007, p. 510).

has parents in N, Z takes its deviant value while all of its parents in N take their default value (and its parents in V\N take their actual values). In this latter case, to explain the deviant value of the child variable, we need to look outside the causal network.

Let's test this formal definition with the Assassination and Bodyguard examples. In Assassination, we can set the default value of C as 0, and the default value of A as 0 as well, since it is not reasonable or natural to expect that someone will poison the drink. The causal network {A, C} connecting A and and C is self-contained: when C takes its default value, its parent A takes its default one as well. More precisely, it is not possible for C to take its deviant value if its parent takes its default one. This matches the intuition that the fact that Alice poisons the drink amounts to a satisfactory and self-sustaining explanation for the death of the victim. What about Bodyguard? Here, the default value of D is set as 0; the default values of A and B are set as 0 as well, as it is not "normal" to expect that Alice will poison the drink, or that someone will administer an antidote.[10] The causal network {A, D} is self-contained, as it is not possible for D not to take its default value if A takes its default one; this, again, matches our intuitions about what counts as a sufficient explanation. {B, D}, on the other hand, is non-self-contained: D can take a deviant value even if B takes its default one. This result matches the intuition that B is not a satisfactory explanation of D.

We have thus met the first *desideratum*: "acts" can be defined as instances of counterfactual causation in a self-contained causal network, like Alice's behaviour in both Assassination and Bodyguard; "omissions" can be defined as instances of counterfactual causation in non-self-contained causal networks, like the bodyguard's behaviour in Bodyguard. In both instances, the outcome counterfactually depends on the agent's behaviour, so both actions and omissions count as causally relevant; the model, however, allows us to discriminate between two types of conduct, or two ways of causing the upshot. This conclusion, arguably, matches our intuitive understanding that Alice "did harm" to the victim, while the bodyguard merely "allowed harm" to occur.

Let's now get to the second *desideratum*, that is, distinguishing omissions from causally unrelated events or, differently put, the selection of causes *versus* background conditions. *Prima facie*, Hitchcock's model seems not particularly well-equipped. Let's take again the Dropping the cigarette example, and test how this model accounts for the connection between the presence of oxygen and the starting of the fire. We can draw here the following causal graph:

[10] Specifically, Hitchcock (p. 507) argues that "temporary actions or events tend to be regarded as deviant outcomes. In the case of human actions, we tend to think of those states requiring voluntary bodily motion as deviants and those compatible with lack of motion as defaults."

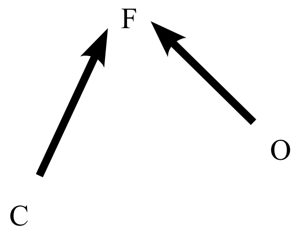


Figure 2.3: Dropping the cigarette

Where:

F = 0 if the fire doesn't start, 1 if it does, and Def(F) = 0;[11]

C = 0 if I don't drop the cigarette, 1 if I do, and Def(C) = 0;

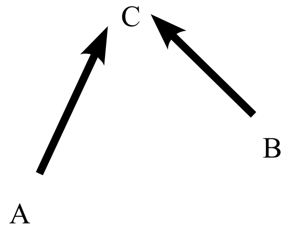
O = 1 if the oxygen is present, 0 if it isn't, and Def(O) = 1;

F = C & O.

The presence of oxygen counts here as allowing harm: F counterfactually depends on O, since the counterfactual "had the oxygen not been present, the fire wouldn't have start" is true. Also, the causal network {O, F} is non-self-contained: the outcome could take its deviant value when O takes its default one. This result, clearly, does not match our intuitive understanding that the presence of oxygen amounts to a background condition, which is weaker than a causally relevant omission.

Yet, according to Hitchcock, we can still successfully distinguish "real" omissions from mere background conditions by appealing to the notions of default and deviant value. Roughly, we can identify whether a variable intuitively counts as "allowing" by looking at the value this variable takes in the causal model: if the variable takes the default value, we can describe it as a background condition, and thus a "false" omission, while if it takes its deviant value we consider it as causally relevant, and thus an allowing.

Let's see how this suggestion works in the Neighbour/Queen of England example. Here, we can build the following causal model:



[11] I refer here to Hitchcock (2007)'s formalisation, where Def(X) for any variable X indicates the default value of the variable.

Figure 2.4: Queen of England

Where:

A = 1 if neighbour waters my flower, 0 if not (where Def(A) = 1);

B = 1 if Queen of England waters the flower, 0 if not (Def(B) = 0);

C = 1 if flowers die, 0 if not, and

C = not-A & not-B.

The fact that A takes 1 as default accounts for the fact that, given the promise the neighbour made, it is reasonable to expect that she does indeed water my flowers. B, on the other hand, takes 0 as default because it is not expected at all that the Queen turns up and waters my flowers. The identification of the causal networks {A, C} and {B, C} self-contained rather than non-self-contained depends here upon which value we assign as default to the outcome C, the flowers dying. I think that the best solution here, while not particularly elegant, is to assign different default values depending on which causal account we are considering. When evaluating {A, C}, Def(C) should be put at 0, since it is to be expected that the flowers die, which is what will "naturally" happen if nobody waters them. The causal network is non-self-contained: the outcome C can take its deviant value 0 (that is, the flowers live) exactly when its parent A takes its default value 1 (the neighbour waters the flowers). This also correctly identifies the neighbour's negligence as a real omission, since not watering the flowers amounts to a "deviant" behaviour. Let's now turn to the case of the Queen. When evaluating {B, C}, we could reasonably put Def(C) as 1, the flowers live, because this is what we expect given that the neighbour was asked to water the flowers. The casual network is again non-self-contained: the outcome can be deviant (that is, the flowers die) even when C takes its default value; this would happen in the case that the neighbour is negligent. The Queen's behaviour also counts as a "background" condition, as her not watering the flowers is the default. This solution, though not particularly elegant, can correctly define the neighbour's behaviour as allowing, but not the Queen's. We can generalise this solution, and thus have that Hitchcock's model is in line with the second *desideratum* too – it can distinguish between background conditions and omissions.

In the following section, I deal with a challenge for most counterfactual accounts of causation, preemption, and I examine whether the self-contained network account can accommodate this difficulty. I conclude that, at least to some extent, preemption cases remain problematic. Nonetheless, as I argue in 2.4.2, this model has some promising advantages, and, despite the difficulties it has in dealing with preemption, it still amounts to an adequate framework of the doing/allowing distinction.

2.4.1 Remaining difficulties: preemption

One issue remains open with respect to this model, namely the technical difficulty of correctly accounting for preemption cases as doings. Recall the standard preemption case of Backup.[12] In this example, Alice poisons the victim's drink, and Backup is ready to do the same if Alice fails her task. The variables in this story thus are:

A = 1 if Alice poisons the drink, 0 if she doesn't;

B = 1 if Backup poisons the drink, 0 if she doesn't;

C = 1 if the Victim dies, 0 if she doesn't.

The equations for this model are:

B = not-A

$C = A \vee B$

The first equation reads as “if Alice hadn't poisoned the coffee, Backup would have”, while the second equation reads as “if Alice or Backup had poisoned the coffee, the victim would have died”. For the sake of simplicity, we can consider the following causal graph:

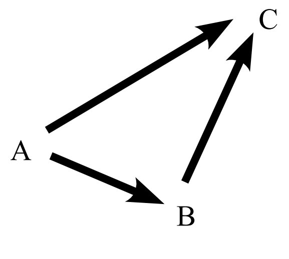


Figure 2.5: Backup

In Backup, {A, C} and {A, B, C} are both paths from A to C. According to Hitchcock's definition, there is only one causal network connecting A to C, which is {A, B, C}. The default value for C is 0, that is the victim does not die. The default value fo A and B is 0 as well, that is, that they do not poison the victim. Despite the fact that Alice's behaviour strikes us as doing harm, the self-contained network model does not deliver here the same result. The causal network{A, B, C}, indeed, is non-self-contained, because B takes its deviant value exactly when its parent, A, takes its default one. Even worse, note that C does not counterfactually depend upon A, as the counterfactual “had Alice not poisoned the drink, the victim wouldn't have died”

[12] Hitchcock (2007), p. 499.

is false. Alice's behaviour thus does not amount to a doing neither an allowing.

Hitchcock (2001) and Halpern and Pearl (2005) provide a promising strategy for restoring the intuition that Alice causes in fact the death of the victim, by the means of further technical refinements. The underlying idea is that in cases of preemption one causal path is "direct", namely {A, C}, while another is "indirect", that is {A, B, C}, as it runs through B. The variable B, Hitchcock argues, makes "some sort of cancellation", which is responsible for the fact that C does not counterfactually depend on A. As Hitchcock puts it "we need to isolate the influence of the former (A) on the latter (the outcome C) along the direct path. We can do this by 'freezing' the indirect path. That is, when we hold the value of B fixed at its actual value of 0, the counterfactual dependence of C upon A is restored."[13] Leaving further technicalities aside, this strategy amounts to consider the counterfactual "if Alice had not put poisoned the drink, and Backup (still) did not put poison in the coffee, then victim would not have died." Intuitively, this counterfactual is true, and the related causal network is self-contained; we can thus conclude that, consistently with our intuitions, Alice did harm the victim. Systematic "freezing" in cases of preemption requires some further technical moves that I do not spell out here in details. Note that, as I anticipated in Chapter 1, accommodating for preemption require us to refine the model introducing aspects that are context-sensitive, like the distinction between direct and indirect causal paths.

This sophisticated model, Hitchcock concedes, is not itself immune to counterexamples: there are in fact even more complex preemption scenarios where our intuitions about an agent's causing an outcome are not captured by the counterfactual model.[14] It is beyond the scope of this section to dwell on these further difficulties. In summary, preemption remains so far not completely resolved in structural equation frameworks, despite the freezing strategy successfully accounts for most simple cases. Regarding more complex ones, I suggest that the self-contained model can bite the bullet, and still serve as an adequate framework of the doing/allowing distinction, in the light of its advantages, which I examine in 2.4.2.

2.4.2 Deviant, default, and norms

Hitchcock's model, I argue, possesses adequate tools to assess the different causal impact of acts and omissions, and to discriminate between real and false omissions, and thus may capture our doing/allowing classifications. The model also makes explicit, in the identification of the

[13] Ibidem, p. 520.

[14] For one of these examples, see p. 520.

deviant or default value of a variable, the role of presuppositions and expectations about what is "normal" in laying the ground for evaluating the agent's impact on the outcome. As Hitchcock argues, in some cases such as Dropping the cigarette, the default value may be set by "objective" or factual features of the context: there usually is oxygen in the atmosphere, and thus the default value for this variable is 1. Arguably, however, in cases involving human agency, the definition of "normal" rather relies on a set of social expectations, including moral principles, obligations, standard rules of conduct and so on. While Hitchcock does not further analyse the process at work in the identification of the deviant/default value of the variables, I think that norm-based accounts can shed light on this mechanism, providing some helpful indications and criteria. In this sense, I suggest that norm-based considerations help "set the scene", in selecting, by the means of fixing the deviant/default value, which conducts amount to normal behaviours, standard explanations or background conditions *versus* actual causes. Ultimately, counterfactual analysis delivers an assessment of the agent's causal impact on the upshot, by the means of identifying relations of counterfactual dependence. The persistent intuition that the way an agent brought about an outcome (act/omission) matters to the doing/allowing distinction is thus captured in counterfactual terms; the insight that what counts as "normal" influences this classification is captured by the norm-based considerations that perform the groundwork.

Let's now discuss in more detail how these norm-based considerations help in setting the value of variables. An appropriate norm-based account, as I argued in section 2.1, should be broadly construed to account for more stringent obligations, such as rights violations, as well as less stringent rules of conduct, such as keeping promises. Kagan's norm violation proposal, in this sense, seems to offer a suitable framework, as it can incorporate the whole set of social expectations which can influence our doing/allowing evaluations. Any specific description of a succession of events, of a choice problem or of a social interaction, commonly encodes and incorporates an understanding of how events should naturally occur, what is expected from the agent in such circumstances and which rules and behaviours are appropriate to the context. Note that, as I will argue at length in the following chapters of this study, some cases are nonetheless under-described; in these examples, it might then be indeterminate what is the normal course of events, and different people might have different expectations about what is more likely to happen. Similarly, some examples may be under-described in yet another respect: it is not clear which are the relevant moral principles or rules we hold each other to in a given context. Ultimately, many of these empirical expectations or moral considerations may possibly apply in a given case. However, when evaluating a specific case, only a subset of moral principles, social expectations or common explanations will stand out as relevant or *salient* to a given agent.

Finally, people might also have different moral convictions and beliefs: they might thus agree on which norms are relevant in a given but disagree on their relative importance, and on which behaviour is, all-things-considered, more appropriate.

Let's start testing this hypothesis with some of the cases discussed in these chapters. In the Neighbour example, what stands out is plausibly the expectation that the person who makes a promise should keep it. In Quinn's trolley case, what seems more relevant is the fact that a person has a negative right not to be run over, or a non-intrusion claim over her body, and thus the agent ought not to violate them. Similarly, a doctor has the deontological duty to perform life-saving treatments; reasonable prudential reasons require that I check my house appliances; the breaking of something, like a window, must usually be explained by an external interfering force, such my throwing a rock; the standard explanation for the occurrence of fire is that something inflammable reacts with the oxygen in the atmosphere; the expected result for a rock's slipping down the hill is that it continues to slip, and so on. Once the commonly expected or agreed upon conduct or succession of events is set, the variable describing the conduct, or event, at issue takes the default value if it complies with these standard expectations and more or less stringent obligations or rules. On the other hand, if the conduct violates one of these rules, or a phenomenon does not occur according to the standard explanation or natural succession of events, the variable describing it takes the deviant value.

This interpretation of norm-violation as a way for setting deviant and default variables in turn allows the classification of a behaviour as a doing or an allowing (or neither). In the Neighbour/Queen of England example, my neighbour seemingly violated the standard rule of behaviour of keeping one's promises; this consideration explains why the variable describing her conduct takes the deviant value 0 in the related causal network model, since, as noted above, expectations set the default value as 1, that is, the neighbour waters my flowers. Furthermore, the default value for the variable describing the flowers dying is 1, as the standard or expected succession of events is for flowers to "naturally" die, unless someone waters them. The standard expectation regarding the Queen's conduct, or other random agents', is that they just go on with their normal life without watering someone else's flowers: in this sense, the Queen watering my flowers would amount to an extremely abnormal behaviour on her part, and thus B takes the default value 0. Once the values are set, Hitchcock's analysis delivers the following verdict: C counterfactually depends upon A, since had the neighbour not watered my flowers, they would have died; the causal network is non-self-contained, since the outcome C sometimes takes its deviant value (the flowers live) when all its parents take their default one (the neighbour waters the flowers and the Queen does not). The relation between B and C is again of counterfactual dependence in a non-self contained network, but C takes here its default value. According to my suggested interpretation of these results, my neighbour's negligence classifies as a causally relevant omission, and thus as“allowing”, while the Queen's action classifies as a false omission, and thus not as “allowing”; this description, seemingly, matches our intuitions about the cases.

This model may successfully accounts for most of our attributions of allowing, such as negligence, refusing to aid or refraining. These are cases where the agent's conduct seemingly violates some kind of normal expectation or standard explanation which is salient for the situation or interaction at issue, and thus the relative value takes its deviant value, *but* the impact of the agent is perceived as an omission, as the upshot was somehow going to occur anyway. On the other hand, norm violations we perceive as doings may have the structure of counterfactual dependence in a self-contained network. Eventually, capturing Kagan's intuition, conducts that do not amount to violation of norms, or alterations in the standard course of events, will be often defined as neither doings nor allowings.

This interpretation of Hitchcock's model, of course, leaves room for disagreements in our doing/allowing classifications. As I argued above, individuals can reasonably have different empirical or moral expectations, depending on which course of events or rule of behaviour they pick as salient. Moreover, as I will discuss in Chapter 4, doing/allowing classifications may as well vary across descriptions. Different “framings” of the apparently same situation or decision problem can trigger the perception of different norms as salient, or explanations as standard, and thus define a behaviour as compliance or violation, or a variable as a background condition rather than a “real” cause. These different expectations, in turn, will deliver different doing/allowing classifications. Eventually, interpersonal disagreement can reflect actual moral disagreement. In Kagan's Respirator example (p. 73), one's moral beliefs about the legitimacy of euthanasia may influence the assignment of default values in the relevant causal network framework, thus inducing a different doing/allowing classification depending on the agent who makes the judgement. I will discuss this example in detail in 5.3.2.

Note too that, in my interpretation of doing/allowing in the the self-contained network model, the “explanation” does not run very deep. Some may argue that, in fact, we are simply putting our intuitive understanding of what counts as doing and what counts as allowing in the value assignment to variables. I am willing to concede this point. For the time being, I have simply argued that this model seems to adequately describe our use of the doing/allowing distinction.

In the next chapters, I will elaborate on these observations. Ultimately, I will argue that accounting for disagreement and context-dependency of doing/allowing classifications amounts to an explanatory advantage of the self-contained network model. Moreover, I will suggest that this model, rather than being “circular”, may help shed light on which normative features the doing/allowing distinction incorporates.

3. The strategy of “fully-equalized cases”

This chapter, together with Chapter 4, sets out to isolate the question of whether the doing/allowing distinction amounts to a morally relevant one. One way to do this, which has been widely employed in the literature, is to look for cases where there seems to be general agreement about the characterisation of actions as doing *vs* allowing, and consider intuitions in these cases about whether the distinction matters morally. I explore here a promising strategy for doing such a test in a controlled manner: the strategy of comparing “fully-equalized cases”.

First, in 3.1, I outline two main positions one could take with respect to the moral relevance of the doing/allowing distinction, which I call the “positive” and the “negative” theses. In 3.2., I introduce three examples of fully-equalized cases, and discuss people's responses to them. In 3.3, I argue that pervasive disagreement, which is a characteristic of discussion on fully-equalized cases, does not seem to amount to evidence in favour of the positive or the negative thesis.

In this chapter, I also argue that when we try to control for all other factors apart from doing/allowing, the examples discussed in the literature become fuzzy, artificial and confusing. Specifically, I claim that when we try to achieve fully-equalized cases, the agreement over how to draw a line between the two different types of conduct and, more importantly, the agreement over whether this distinction is morally relevant might start to look more tenuous. In some examples of fully-equalized cases, disputes about whether the distinction matters morally are intimately tied up with disputes about how to classify actions as doing *vs* allowing in the first place. Other examples, in the attempt to separate doing/allowing from, for instance, the intending/foreseeing distinction, end up being extremely contrived, such that intuitions about these examples do not seem reliable. As a result, I conclude that moral reasoning over fully-equalized cases is not a good guide to moral truths. Anecdotal evidence and some problematic features of empirical surveys in this field (3.3.3) also suggest that the underlying rationale of the fully-equalized cases strategy – disentangling the doing/allowing distinction from any other morally relevant feature of the context – might prove unfeasible.

3.1 The question of moral relevance

In the previous chapters, I outlined and compared alternative accounts of the doing/allowing distinction, assessing their success in distinguishing between doing and allowing behaviours. The focus of this task, in short, was to identify which framework is descriptively adequate for capturing our intuitive and everyday use of the distinction. My interpretation of Hitchcock's model, I argued, best fits the way we distinguish between doing *versus* allowing behaviours.

While the issue of the descriptive adequacy of moral categories is an interesting analysis *per se*, its traditional role in moral philosophy is that of a preliminary enquiry and necessary tool for dealing with the more significant question of moral relevance. This point is particularly true when it comes to the doing/allowing distinction. As I argued in Chapter 1, the principle that "doing harm is morally worse than allowing it to occur" strikes us as uncontroversial and obvious in very many cases. In moral philosophy, this principle has been famously defended by, among others, Frances Kamm (1996) and, more recently, Fiona Woollard. As Woollard (2015)[1] puts it, in fact, it seems that:

> "If there is no moral difference between doing and allowing, then morality must either be far more permissive than we generally suppose – permitting us to kill to protect our personal projects – or far more demanding – requiring constant sacrifice from us to save the lives of others".

Despite the persistence and reasonableness of this intuition, however, there is still disagreement in the literature over the *true* moral relevance of the doing/allowing distinction, once we get down to the details. Opponents of this principle, that doing harm is worse than allowing it, such as Walter Sinnott-Armstrong (2008), James Rachels (1975), Shelly Kagan (1989) and Jonathan Bennett (1995), claim that the mere fact that a behaviour is an instance of doing, and another is an instance of allowing, is not morally relevant in judging the former to be worse than the latter, offering a wide range of counterarguments.[2]

In this chapter, I will argue that the failure of the strategy of fully-equalized cases seems to show that the separation between the descriptive question and the question about moral relevance might be harder to achieve that one might hope. This conclusion, while not necessarily ruling out the position that the doing/allowing distinction is morally relevant, might challenge the

[1] Fiona Woollard (2015) pp. 17–20.

[2] Recall that, in this study, I discuss the issue of the moral relevance of the doing/allowing distinction referring to the distinction between doing *harm* (or doing something bad) and allowing *harm* (or allowing something bad to happen). Of course, the moral relevance of doing/allowing could also be discussed by referring to doing good/allowing something good to happen, and to the moral principle that "doing good is morally better/preferable/more praiseworthy than allowing the same good to occur". The first interpretation, however, is far more discussed in the literature; it is not straightforward, moreover, that these two moral principles should be simply treated as symmetrical. In this sense, "doing good is better than allowing it" would require a separate and specific analysis. In the remainder of this thesis, therefore, whenever I talk about the moral significance of the doing/allowing distinction, I refer to the "harm" interpretation.

apparently intuitive idea that doing and allowing are morally significant features *per se*. In chapter 4, I will continue this task and focus on what I perceive to be the second main challenge to the moral relevance of the doing/allowing distinction: frame-dependency.

In this chapter, I do not survey all of the different accounts which have been upheld in the literature, both by advocates and opponents of the moral relevance of the doing/allowing distinction, but rather I outline two main positions, with respect to whether doing and allowing amount to morally relevant features.

I summarize the two main positions concerning the moral relevance of the doing/allowing distinction as follows:

> **Positive thesis**: there is a fact of the matter whether a behaviour is an instance of doing rather than allowing; this distinction is morally relevant as, *all other things being equal*, a "doing" behaviour is morally worse than an "allowing" behaviour.
>
> **Negative thesis**: there might, or might not, be a fact of the matter whether a behaviour is an instance of doing rather than allowing. *Whichever the case*, this distinction is not morally relevant for assessing the relative moral value of actions.

Let's clarify the implications of the positive and negative theses, as I formulate them above.[3] First, note that, the way the positive thesis is characterised, there could still be room for a "doing-harm" behaviour to be morally preferable to an "allowing-harm" behaviour, if other considerations are more significant in the moral evaluation of cases. In other words, the positive thesis does not require that doing harm is always morally worse than allowing a harm to occur, independently of any other consideration; the thesis rather makes the much more limited claim that this distinction is morally relevant. Secondly, my loose formulation of the negative thesis does not differentiate between positions which deny that doing and allowing behaviours can be descriptively distinguished, and accounts which allow for this empirical distinction but still argue that it is not morally significant. Either way, the negative thesis, so formulated, states that the doing/allowing distinction has no distinct role in assessing the moral value of actions.

It is also worth clarifying the relationship between a) descriptive frameworks of the doing/allowing distinction, which may be more or less adequate in capturing our everyday use of these classifications, and b) different accounts of the moral relevance of the distinction. As I argued in Chapter 2, my proposal is that a "mixed" account, like Hitchcock's, can quite successfully track our doing/allowing attributions in most scenarios. This conclusion, however, is not necessarily evidence in favour of the position that the doing/allowing distinction is

[3] This characterisation of the positive and the negative theses does not refer to any particular framework or account of the doing/allowing distinction.

morally relevant *per se*, or even that there is a fact of the matter whether there are "correct" doing/allowing characterisations. The self-contained network model, in this sense, merely accounts for our ordinary *use* and intuitions about the distinction. In this chapter, I will thus assume that the descriptive problem has been solved. Therefore a background assumption is that doing/allowing is best cashed out in terms of Hitchcock's model, but in fact, the arguments in the chapter do not make reference to this model.

Here, in particular, I outline and discuss a common strategy, employed by both the supporters of the positive and negative theses, for assessing whether the doing/allowing distinction is morally significant *per se*, once all other features of the behaviours at issue are controlled for. In 3.2 I simply describe the strategy, including key examples in the literature and responses to these examples. In 3.3, I will then turn to a more in-depth analysis of the viability of the strategy.

3.2 Discussing fully-equalized cases

First, I should address a naïve line of argument against the positive thesis. Compare, for instance, our moral judgements in the two following actions: Killing an assailant in self-defence and Refusing to rescue a person tied to a track. Reasonably, most people would classify the former action as an instance of doing harm, and the latter as merely allowing a harm to occur, as it boils down to refraining from untying the person. There is no disagreement, in short, over the factual description of the two behaviours. Nonetheless, contrary to what supporters of the positive thesis claim, most agents also argue that the latter is morally worse than the former, or, more specifically, that Killing an assailant would be morally permissible, while leaving the person tied would be morally wrong. This, apparently, violates the intuition that "doing is worse than allowing".

Clearly, however, this counterexample is extremely naïve and does not effectively challenge the positive thesis: for doing and allowing to amount to morally relevant features, it need not be the case that all instances of doing harm are worse than all instances of allowing harm, no matter the circumstances. This point is true for many other moral principles. While conceding that, for instance, loyalty is morally better than disloyalty, we can grant that committing a murder for the sake of loyalty is, all things considered, morally bad. In other words, the positive thesis only claims that "doing harm is worse than allowing it" *all other things being equal*. For the "doing" and "allowing" characterisations to be morally significant, therefore, we do not need to prove that they alone and univocally determine our moral judgements in all cases. In Killing an assailant *vs* Refusing to rescue the tied person, for instance, other moral considerations, like the right to defend oneself and the duty to rescue, play a larger role in the evaluation of cases.

In short, it can be argued that Killing an assailant and Refusing to rescue are vastly different cases, and that their comparison is not a reliable indicator of the role of the doing/allowing distinction in moral judgements. Too many independent considerations affect the relative assessment of these behaviours. If we seek to analyse the specific import and role of the doing/allowing distinction, therefore, we must make sure that *only and exactly this* feature decides our relative moral evaluation of cases. To have strong evidence in favour of the positive thesis and against the negative thesis, therefore, we need to show that when two cases share all the morally relevant features, and only differ in the fact that one is an instance of doing a harm, while the other amounts to allowing the harm to occur, the former is judged as morally worse than the latter.

What has just been described amounts to the strategy of so-called "fully-equalized cases". This strategy has been employed and defended, among others, by Jeff McMahan (2013), who argues that this amounts to an epistemically reliable method to test our moral intuitions, "filter(ing) out irrelevant details which could distract and confuse them, thereby allowing us to focus on precisely those considerations that we wish to test for moral significance.".[4]

Note that the strategy of fully-equalized cases seems to require a consensus on what actually amounts to fully-equalized cases. I will examine this issue more thoroughly in 3.3. For the remainder of 3.2, I put this question aside and describe three influential examples of (allegedly) fully-equalized cases. Specifically, I first describe such cases (3.2.1) and discuss responses in these examples (3.2.2).

3.2.1 Candidate fully-equalized cases

In this section, I describe three influential examples of fully-equalized cases: the Smith/Jones example by Rachels (1975), Thomson's (1986) trolley examples, and the active/passive euthanasia case.

The Smith/Jones example

The first, most influential, pair of examples is from Rachels (1975, pp. 78–80), who describes

[4] McMahan (2013), p. 9.

the two following (allegedly) fully-equalized cases:

> **Smith (doing harm)**. Smith drowns his cousin in the bathtub in order to inherit a large sum of money.
>
> **Jones (allowing harm)**. Jones, motivated by the same intention and with the same plan in mind, finds his cousin already drowning in the bathtub and refrains from saving him, watching him die.

The two cases, Rachels argues, are identical except for the fact that one is commonly classified as doing a harm, while the other can be thought of as an instance of allowing a harm. This classification clearly complies with most descriptive frameworks discussed in Chapters 1 and 2. He then claims that Smith's and Jones's behaviours are morally equivalent, and that, consequently, “doing” is no worse than “allowing”, when we get rid of all other distinguishing factors. He argues that the reason this distinction is thought of as morally relevant is that cases of doing harm are usually “crueller”, or more callous, due to other features that set them apart from allowings, which are properly controlled for in this example.[5]

Specifically, Rachels' case seems well-controlled because both Smith and Jones have the same *intention* when entering the bathroom: to kill their respective cousins. I will come back later (in 3.2.2) to the question of whether we can really control for intentions; for the moment, I just note here that both Smith and Jones desire as the outcome of their action the death of their cousin. This is the goal both individuals' conduct aims at, rather than the death of the cousin being a (more or less regrettable) side effect of their acting. This is a particularly clever aspect of the example, since most cases of allowing harm are effectively cases of *foreseeing* rather than *intending* harm in the sense just alluded to. Roughly, I am foreseeing a harm when I can expect or reckon it to be a side-effect of my acting, but the real aim of my action is a distinct and separate one.[6] For instance, when Refraining from rescuing the tied-up person, which is a clear-cut case of allowing, we might think that killing that person is not within the realm of the agent's aims and goals, and so the harm done is not intended. In Rachel's case, however, the “allowing” behaviour is arguably motivated by the same intentions as the “doing” behaviour.

The trolley example(s)

Philippa Foot (1978) and Judith Jarvis Thomson (1986) provide a second influential model of

[5] Rachels (1975), p. 16.

[6] In the present chapter, I will only talk about intentions in an informal way, appealing to our intuitive understanding of this notion. In Chapter 6, I will discuss the intending/foreseeing distinction more thoroughly.

the strategy, comparing two fully-equalized cases involving the choice between the death of one or five innocent people. To be clear, neither Foot nor Thomson define theses examples as fully equalized cases, but rather use them as a way to argue for or against the permissibility of killing. In doing so, however, the doing/allowing distinction also seems to become relevant, and for this reason trolley cases are often discussed in relation to the debate. As it will emerge from my discussion here and in chapter 6, I do not think trolley cases pose a threat to most accounts of the doing/allowing distinction. The reason why I discuss trolley cases here is that some *pairs* of trolley cases could be (in my view, wrongly) described as equalized cases, which only differ in the fact that one compares doing with doing (or allowing with allowing) and the other doing with allowing.

To place trolley cases in the literature, Foot argues that doing harm (or, using her terminology, killing) is worse than allowing harm (or letting die) because killing one person is worse than letting five die, and people would choose the former over the latter. On the other hand, if we had to choose between killing one and killing five (or letting one die and letting five die), we might choose the action which minimizes the number of deaths.[7]

Thomson challenges this view by arguing, on the contrary, that we can build two fully-equalized cases (again, this terminology is not Thomson's), both involving the choice between killing one and letting five die, where our intuitions about the permissibility of killing one change. If this were the case, we would have a case against the intuition that, all other things being equal, killing (doing) is worse than letting die (allowing). To do so, Thomson compares different cases. I focus here on Fat Man *versus* Bystander.

Fat Man:

> “you are standing on a footbridge over a trolley track. You can see a trolley hurtling down the track, out of control. You turn around to see where the trolley is headed, and there are five workmen on the track where it exits from under the footbridge. What to do? Being an expert on trolleys, you know of one certain way to stop an out-of-control trolley: Drop a really heavy weight in its path. But where to find one? It just so happens that standing next to you on the footbridge is a fat man, a really fat man. He is leaning over the railing, watching the trolley; all you have to do is to give him a little shove, and

[7] To illustrate this point, Foot uses herself a Trolley case, which is identical to Thomson's Bystander, but it is the *driver* who chooses between switch and don't switch. Trolley is then compared with Riot: you can either execute an innocent man (killing), or let five men die in the riot which will inevitably follow if you don't execute the one man. Foot then argues that, in Trolley, both Switch and Don't Switch, if performed by the driver, amount to killing. While it will be impermissible to execute in Riot, it seems permissible to switch in Trolley. Therefore, Foot concludes, agents are appealing to the principle “killing is worse than letting die”, as they would minimize the number of deaths only if that does not require choosing the “doing” conduct when the “allowing” conduct is available.

> over the railing he will go, onto the track in the path of the trolley.'[8]

According to Thomson, in Fat Man, the available options are:

> **Push (doing harm)**. Push the innocent fat man and stop the trolley.
>
> **Do nothing (allowing harm)**. Let the trolley run over five innocent workmen.

Thomson argues that, in this case, it is impermissible to Push, so it seems that killing one is morally worse than letting five die.

In Bystander, Thomson describes a different trolley scenario:

> The trolley is running towards five track workmen. There is a spur of track leading off to the right. On the spur of the track there is another workman. You are a bystander observing the situation, and you could throw the switch, thus turning the trolley.

Here, Thomson argues that the two options are characterised as follows:

> **Don't switch (allowing harm)**. Let the trolley continue along its track and run over five innocent people.
>
> **Switch (doing harm)**. Pull the switch and turn the trolley which runs over one innocent person.

Here, Thomson argues that it is permissible to turn the trolley, and thus, killing one is *not* morally worse than letting five die.

The two cases, according to Thomson, are identical in all respects, as they both involve a comparison between doing harm to one and allowing harm to five. *Contra* Foot, however, Thomson concludes that people do not seem to appeal to a moral difference between doing and allowing, as this would not explain people's different intuitions in these two cases. Arguably, we would need to appeal to other features of the two cases so as to justify different moral judgements. For instance, it could be argued that the agent's intentions in Fat Man are not the same as in Bystander. It is true that, in both cases, the ultimate goal is to save the five workmen; nonetheless, at least according to some accounts of intentionality, we might say that when one throws the switch one is merely foreseeing the death of the person on the spur of the track, while when one pushes a man onto the track one intends the death of the man – even if this is a means to save five. I will come back to this issue in 3.3 and, more thoroughly, in Chapter 6 of this thesis.

Also note that it is at least controversial whether, in Bystander, Switch is an instance of doing harm. On Foot's account, for instance, both Switch and Don't Switch amount to allowing, as

[8] Thomson (1986), p. 1409.

neither of them involves initiating or sustaining a harmful sequence. Among others, Rickless (1997) also argues in this direction. Again, I will discuss this lack of consensus in 3.3.

The active/passive euthanasia example

The debate over the permissibility of euthanasia often revolves around the comparison and relative evaluation of different medical practices in end-of-life situations. The standard way[9] of distinguishing between and legislating on euthanasia practices, in legal codes, bioethics and common discourse thus usually draws a line between so-called:

> **Active euthanasia (doing)**. Behaviours like administering a lethal injection.
>
> **Passive euthanasia (allowing)**. Behaviours like refusing to provide life-sustaining treatments, switching off of life-supporting machines and so on.

This distinction seemingly tracks people's, including medical professionals', persistent intuition that injecting a lethal drug into a dying and suffering patient would amount to killing her (doing harm), while merely denying treatment, or even "unplugging the machine", should be classified as letting die (allowing harm). In both situations, however, the intentions of the doctor performing the practice are to bring about the death of the patient so as to end her pain and suffering; moreover, we may assume that both behaviours are equally effective in bringing about the outcome of the death of the patient. Thus, these two cases are apparently fully-equalized, such that the only difference is that one is a doing and the other an allowing. If we do perceive the two actions to be morally distinguishable, therefore, this seems to be evidence in favour of the positive thesis.

3.2.2 Responses to the fully-equalized cases

In this section, I briefly discuss responses to the three fully-equalized cases above, to see whether these examples provide strong evidence in favour of either the positive or the negative thesis. I conclude that none of them does, as there is no shared intuition in these cases. This persistent disagreement gives us reason to examine whether these cases are actually controlled, and, if they are, whether controlled cases are indeed a reliable test for our intuitions over the moral relevance of the doing/allowing distinction.

In Smith/Jones, Rachels argues that ultimately the "doing" and the "allowing" conducts are

[9] NHS, https://www.nhs.uk/conditions/euthanasia-and-assisted-suicide/.

morally equivalent. Contrary to this intuition, however, it can be argued that it is at least disputable that Smith's action is not, all things considered, worse that Jones's. In criminal law, for example, Smith would surely be regarded as guiltier than Jones. Frances Kamm (2007) further argues that the two cases are not intuitively equivalent, as we could not impose the same losses on Smith and Jones, if those losses were necessary to bring the cousin back to life: specifically, she claims that it would be permissible to shoot Smith, but not Jones, if this would be necessary to bring the cousin back.[10]

The trolley cases, even more ostensibly, are the subject of ongoing debate. It is extremely controversial whether it would be permissible to Switch in Bystander, as well as whether Switch amounts to doing harm or allowing harm. Ultimately, disagreement is also the trademark of the active/passive euthanasia debate. On the one hand, this distinction, which arguably tracks the doing/allowing dichotomy, and thus different ways of being relevant to the harm, is often intuitively recognized and employed for legislative purposes. On the other hand, many scholars argue that the two conducts are morally equivalent and the active/passive distinction is tenuous, irrelevant or obscure.[11] Here, again, we thus seem to see disagreement over the moral relevance of the doing/allowing distinction.

In all three examples, the most evident result seems, therefore, to be an overall and persisting lack of consensus – both in the literature and in people's intuitive responses –[12] regarding the assessment of otherwise equalized cases. The idea that we can remove all confounding factors, and then see whether doing and allowing behaviours are, or are not, morally equivalent is thus at least controversial. However, some might still consider the disagreement as some evidence for the positive/negative thesis. At first sight, the lack of consensus could be seen as bad news for the positive thesis. After all, if doing and allowing were morally significant characteristics, it should be expected that, when properly isolated, their contribution would strike most people and authors as clear-cut and straightforward, thus leading to a smaller number of “deviations” and a more robust agreement across individuals.

This, however, would be a premature conclusion. In fact, the lack of consensus may equally be used by advocates of the positive thesis to argue that there is at least some resistance to assimilating doing and allowing behaviours, thus challenging the negative thesis.

It seems, therefore that little can be concluded from the fully-equalized cases strategy. Lack of consensus and persistent disagreement could in fact be symptomatic of a number of things. It

[10] Kamm (2007), p. 17.

[11] Some notable examples in moral philosophy are Rachels (1975), Dworkin, Nagel, Kamm and Thomson (1997), who discuss the matter thoroughly in “The Philosophers' Brief”.

[12] For some experimental surveys, testing subjects' intuitions in different trolley cases, see Petrinovich, O’Neill and Jorgensen (1993); Greene *et al*. (2001), Greene (2015).

could be that the doing/allowing distinction does not matter morally, or that it does matter but there is disagreement over the analysis or detection of a doing *versus* an allowing conduct. In the next section, I examine more closely the nature and the reasons for this persistent disagreement. My conclusion is rather negative: the strategy of fully-equalized cases is not successful in illuminating moral reasoning.

3.3 Problems with the strategy

The fully-equalized cases strategy, as I argued in 3.2, has some undeniable merits and seemingly amounts to a promising tool for isolating the role of the doing/allowing distinction in our moral evaluations. My contention, however, is that the idea of eliciting reliable intuitions regarding fully-equalized cases with respect to the doing/allowing distinction is ultimately untenable. I argue that, in general, one of two problems arise. First, fully-equalized cases cannot be actually achieved, whether because not all factors are controlled for (trolley examples), or because in controlling for other factors, we also lose consensus over factual doing/allowing characterisations (trolley and euthanasia). The second problem is that even though fully-equalized cases are actually achieved, they come at the cost of the examples becoming so contrived and confusing that they compromise intuitions (Smith/Jones example). In this section, I will discuss how these problems inevitably arise in the process of trying to equalize cases with respect to doing/allowing. I start with equalizing consequences (3.3.1), which is relatively straightforward. Things start to unravel, however, when we try to equalize for intentions (3.3.2). In 3.3.3, I conclude with some remarks on the methodology of testing people's responses to fully-equalized cases.

3.3.1 Controlling for consequences

First of all, it is worth examining more carefully what Rachels, Foot, and Thomson, mean by “equalized cases”, with respect to the doing/allowing distinction. In particular, we can formulate this question as: what particular moral features do they think must be controlled for? The first of these features is the impact or seriousness of *consequences*; the “harm” brought about by the compared actions must be equivalent.

What we need to clear from the picture is the fact that, in the majority of cases, behaviours which are classified as "doing harm" have worse consequences than behaviours we commonly describe as "allowing". Generally, it can be argued that, if I drown a person (a typical case of doing harm), the outcome of my action will almost certainly be her death, while if I refrain from giving money to a beggar (a typical case of allowing harm), the chances that she will eventually die are much lower. Generally speaking, it is frequently observed that cases in which I merely fail to provide aid involve far less serious expected consequences than cases where I bring about a harm or, in other words, it may appear that "allowing" behaviours merely raises the probability of harmful consequences occurring, while the "doing" actions bring them about for sure, or at least the weaker claim: the probability of harm in the doing cases is higher than in the allowing cases.[13]

This consideration, to be sure, could initially raise some questions over the stance of the doing/allowing distinction. It could be claimed that there is nothing more to the doing/allowing distinction than the fact that instances of "doing" have worse consequences than instances of "allowing". It is not difficult, however, to come up with examples where allowing behaviours have no better outcomes than doing behaviours. For instance, the outcome of the action "drowning a person" (doing harm), is equivalent to the outcome of the action "refraining from saving a person from drowning" (allowing harm), when nobody else could possibly perform the rescue, and where the conditions are such that the person is virtually certain to die unless aided. When we equalize outcomes, we see that the doing/allowing distinction does not simply amount to the relative seriousness or probability of negative consequences.

Furthermore, consequences-equalized cases, such as the Drowning *vs* Rescuing example, do not appear particularly unrealistic or artificial. Indeed, there are quite natural cases which even reverse the usual pattern of (likelihood of) bad consequences, such that the allowing behaviour has worse consequences than the doing behaviour. This is true of the trolley examples discussed above. If the cases are equalized in all other ways, this can be useful for making the relative badness of doing, *per se*, more pronounced – it is bad enough to outweigh the consequences being more benign. But there is a worry that trolley cases are not equalized in other ways with respect to intentions, as I briefly mentioned in 3.3.1, and as I discuss in more detail in the section below.

[13] Causal accounts taking the general counterfactual form "E causes F iff had E not occurred, F would have been less likely" could for instance draw a line based on "likelihood" of harmful consequences occurring.

3.3.2 Can we *really* control for intentions?

In this section, I show that the problems of the strategy of fully-equalized cases start when we try to control for intentions. I discuss this claim by examining each of the three cases in turn. I start with the Smith/Jones case, which is the one which seeks to equalize intentions more straightforwardly.

The Smith/Jones example

The Smith/Jones example not only equalizes consequences, but appears to also equalize intentions. In fact, Rachels makes a point of this. He argues that the *intentions* of Smith and Jones must also be the same – killing the cousin in order to inherit – if we want to assess our intuitions about doing and allowing *only*. Rachels' worry here is to make sure that the doing/allowing distinction does not overlap with another widely recognized dichotomy in moral theory, the one between intending and foreseeing. This is not the place here to discuss this moral categorization at length; I thus stick to an intuitive understanding of what counts as intending harm. The intending/foreseeing distinction clearly strikes us as morally relevant: intuitively, intending a harm seems morally worse than merely foreseeing it. Furthermore, this principle appears to have such fundamental moral significance that it could be questioned whether the doing/allowing distinction actually amounts to a separate and independent moral characterisation.[14] In many cases, such as Drowning *vs* Rescuing, these two distinctions seem, in fact, to collapse. The consequences of not aiding a person drowning (allowing) and drowning a person (doing) may indeed be the same, that is, a dead person, but while in the former case our actions can be described as merely foreseeing the harm our action brings about, in the latter the action can be described as intending this harm directly. We could argue, therefore, that the reason why drowning seems much worse than not rescuing captures this aspect, rather than the doing/allowing distinction.

The Smith/Jones example seems to do a better job, and the relative simplicity of this pair of cases seems to be good news. Nonetheless, as I argued in 3.2.1, what should be concluded from the Smith/Jones case is controversial. While Rachels, Sunstein (2003) and other authors argue

[14] In Chapter 6 of this thesis, I will discuss in detail the relationship between the doing/allowing and the intending/foreseeing distinction. I will argue that the two distinctions do not overlap, but doing harm may capture, in some specific circumstances, our judgements on whether the agent intended the outcome. For the moment, however, I will simply assume that the two distinctions are separate, whilst illustrating the difficulties in disentangling the two.

that Jones's behaviour is no better than Smith's, Trammell (1975, 1979) and, more recently, Kamm, claim that Smith's behaviour is more objectionable. I argue that the lack of consensus may be caused by two distinct reasons. On one hand, there might be substantial disagreement over the fact that doing and allowing amount to morally relevant features of actions, when outcomes and intentions are controlled for. If this were the case, the Smith/Jones example could actually amount to a test of either the positive or the negative thesis. Unfortunately, this test does not seem to produce the agreement one would hope for. My contention, however, is that disagreement may also depend on the fact that this scenario is particularly difficult, and puts us in a cognitively stressful situation, in which it is reasonable to expect that people's intuitions are not very clear-cut and precise.

To see this point, let's focus on Jones's story. In this example, Jones has a specific goal in mind, that is, the death of his cousin, and a precise plan to achieve his goal, that is, drowning the cousin in the bathtub. We find out that, for what seems to be an extremely lucky – at least for Jones – coincidence, the plan is already unfolding: the cousin is drowning by chance without him doing the drowning. This turn in the story, I think, makes this scenario very contrived and confusing. While we experience luck and coincidence in our everyday lives, it also seems that in this case luck plays a crucial part – after all, it is by mere chance that Jones does not end up doing harm! I conclude that testing people's reactions in this particularly tricky and artificial case might not provoke intuitions which are reliable, or, most importantly, which can be easily generalized to more familiar cases.

In summary, the Smith/Jones example seems indeed to control for intentions. Nonetheless, there is by no means agreement over whether Smith's behaviour is worse than Jones's. Moreover, equalizing intentions appears to create here a contrived case.

The active/passive euthanasia example

This example also seems to successfully equalize for intentions – in both the active and the passive case, the intention is to bring about the death of the patient, thus ending pain. In fact, it has been suggested (Rachels 1975) that the Smith/Jones example – which seems indeed fully-equalized – is exactly equivalent to the active/passive euthanasia case. Nonetheless it is controversial whether active euthanasia is indeed morally worse than passive euthanasia, or whether the two classes of practices are morally equivalent. I argue here that, in this case, contrary to my original assumption throughout this chapter, disagreement over whether the doing/allowing distinction matters morally is ultimately tied up with disagreement over where to draw the line between doing and allowing. Hence, the descriptive problem and the moral relevance problem may be harder to separate than one might hope.

My claim is that people could disagree in the first place on whether one specific medical practice, such as disconnecting a feeding tube, removing a respirator, or even administering a lethal injection, amounts to an instance of killing (doing harm) or letting die (allowing harm). This descriptive disagreement cannot be put aside when examining whether doing is worse than allowing. I have suggested that the descriptive question of how to distinguish "doing" from "allowing" behaviours can be solved by appealing to a suitable causal model such as the self-contained network account. Nonetheless, there still seem to be "borderline" cases in which doing/allowing classifications might be ambiguous. This hypothesis will be the central claim of Chapters 5 and 6 of this thesis. For the moment, note that the euthanasia example differs in at least one respect from the Smith/Jones case: while the cousin in Smith/Jones is alive and well, and would not die if he did not drown in the bathtub, the practices under investigation here apply to end-of-life situations.[15] We might thus think that, within a reasonably short period of time, the patient would die anyway. The doctor's conduct, in this sense, seems to hasten death, which will inevitably (and arguably soon) occur. To be sure, the doctor is somehow relevant to death of the patient, and often seems to perform a specific practice, and this is ultimately the reason why we talk about this example in relation to the doing/allowing distinction. Nonetheless, when it comes to evaluating the type of impact the doctor had on the death of the patient, it looks like our answers could be different depending on which specific description of the outcome we use. Specifically, if the outcome is broadly described as "death of the patient", we could think that the doctor had an impact on how the death was brought about, but ultimately he "did not" cause the outcome. On the other hand, if the outcome is described as "death at specific time *t,* in this specific way", we could legitimately argue that the doctor "did" cause the outcome. In Chapter 5, I suggest a more thorough analysis of this example using the self-contained network model. Even prior to formal analysis, however, both classifications seem tenable and justifiable.[16]

If my hypothesis is correct, then in the euthanasia case we would have two cases which are actually fully-equalized, but where the characterisation of one conduct as doing and the other as allowing is not uncontroversial or straightforward. Therefore, this example does not amount to reliable evidence in favour of either the positive or the negative thesis. Even worse, this example would show that the task of investigating the question of moral relevance

[15] This is, of course, a simplification, as we could also speak of euthanasia in situations where the patient is not terminally ill. Here, for the sake of the argument, I exclude this case.

[16] This disagreement seems to be supported by empirical results. Cushman, Knobe and Sinnott-Armstrong (2008), for instance, performed a survey with borderline euthanasia examples, asking subjects to classify them as doings or allowings. The results show a significant disagreement over the correct characterisations (even prior to the authors' introduction of "morally ambiguous" descriptions).

independently of the descriptive problem is ultimately unsuccessful, at least in borderline cases.

The trolley example

Let's finally examine the nature of disagreement in the trolley examples, and see whether they actually control for intentions. Unlike the two previous cases, it is controversial here that Fat Man and Bystander are properly equalized. As I briefly mentioned above, one might argue that Push, as opposed to Do Nothing, Switch, and Don't Switch, cannot count as merely foreseeing harm since it amounts to using a person as a means to an end. This difference could explain our intuitions that Push is not permissible, as opposed to Switch. If this were the case, the Fat Man *vs* Bystander case would thus simply not control for intentions, at least within some interpretations of intending/foreseeing.

For the sake of the discussion, let's nonetheless assume that all four conducts are instances of foreseeing the death of one person, as a side-effect of saving five. First, as Thomson (2008) has recently argued, it is at least controversial that, in Bystander, Switch is permissible. Second, it is also controversial whether Switch counts as doing or allowing. After all, one might say, in contrast to Foot and Thomson, it is the trolley which ultimately runs over the one person. Again, if this were the case, we would not have evidence in favour of the positive thesis nor the negative thesis. Like for the euthanasia examples, the conclusion would thus be that what looked to be straightforward cases in terms of classifying actions as doing/allowing, are not so on closer inspection. Finally, I still think that the trolley cases do not adequately isolate the import of the doing/allowing distinction, and thus cannot serve as a test for its moral relevance, even if we grant that intentions are properly controlled. Note that the trolley cases have the structure of “moral dilemmas”; that is, it looks as if none of the options is clearly morally required, yet agents are forced to make a choice.[17] Agents, in short, are being asked how far they would go (or, more precisely, how far they think it is permissible to go) in order to save five people. Among all the different considerations they could ponder, there is *also* the fact that the options might differ in terms of doing/allowing classifications. What we are ultimately testing here, in conclusion, would be the import of the doing/allowing distinction on such evaluation. I think this amounts to a very convoluted way to test for the moral relevance of the doing/allowing distinction.

In conclusion, none of the examples discussed in 3.2 seems to provide a decisive case in favour of either the positive or the negative thesis. To start with, it is controversial whether the trolley

[17] Thomson (1986) makes this point arguing that in trolley cases most people would answer, for instance, that Switch is, all things considered, permissible.

cases adequately control for intentions. Even if they do so, in trolley cases and euthanasia examples, there is further controversy over the factual characterisations of behaviours as doings or allowings. This observation gives us reason to be sceptical that, at least for "borderline" cases, we can successfully keep the factual and the moral issues apart. Finally, in the Smith/Jones case, which seems actually fully-equalized, there is still disagreement over the equivalence of the doing and the allowing behaviours. In the next and final section, I will further explore my earlier remarks about the difficulty and artificiality of fully-equalized cases.

3.3.3 Testing intuitions in fully-equalized cases

From my survey above, two hypotheses remain open with respect to the source and nature of disagreement in fully-equalized cases. It could be the case that the controversy reflects substantial disagreement over whether doing is worse than allowing, that is, between supporters of the negative and the positive theses. However, disagreement could also point to the fact that this strategy is not particularly adequate in eliciting our moral intuitions. In this last section, I discuss some remarks in favour of the latter option. To be clear, I still think that the two hypotheses can be true at the same time; I, thus, just emphasize here the importance of the latter.

Firstly, when presented with Rachels-style examples, it is not uncommon to find a strong tendency from individuals of various and disparate backgrounds to refuse even to engage with these exercises, as well as a reluctance to provide decisive answers. I observed these kinds of responses and attitudes to be almost equally predominant among philosophy-educated subjects.[18] Most students in my classes, fellow researchers and people from different fields made explicit their discomfort in thinking about Smith/Jones, and looked doubtful about the possibility of reaching a reasonable and meaningful resolution. This persistent reaction seems to reflect my contention that the attempt to achieve fully-equalized cases appears to disable our moral intuitions and to challenge the common strategies we standardly employ in moral reasoning.

These kinds of common reaction that I witnessed while researching this topic lead me to question the role that empirical surveys play in this debate. After all, anecdotal evidence amounts to a low-probative type of empirical finding, and more structured experimental results

[18] I would like to thank for this paragraph all my fellow PhD colleagues, who were thoroughly and repeatedly surveyed on these cases. I am also grateful to my audience at King's College Graduate workshops.

over fully-equalized cases are available in the literature. I think, however, most experimental settings may not adequately test for the moral relevance of the doing/allowing distinction; specifically, I argue, they do not help to illuminate whether empirical disagreement actually points to substantial controversy over the moral relevance of the distinction or rather to the fact that fully-equalized cases appear puzzling and artificial, challenging subjects' intuitions and reasoning skills.

But what would it actually take to test these for these two hypotheses? A way to do so would be to systematically keep track of the kinds of reactions expressed above (refusal to engage, confusion), and to ask subjects to motivate their answers. Experimental surveys, however, do not always include as available options "I do not know", or "I do not think I can meaningfully solve this task".[19] When they do so, the prevalence of these answers is not specifically investigated. Greene et al. (2009), for instance, test different trolley cases, and simply exclude from analysis subject "who reported being unable/unwilling to suspend disbelief (31). (...) as well as data from 10 subjects reporting confusion."[20] Note that the survey was performed on 664 subjects, so people who did not engage are more than 6% of the total. What would be helpful, moreover, would be also reporting qualitative evidence about how individuals perform their value judgements, or their reaction to the task. Some notable exceptions, such as Jou, Shanteau and Harris (1996) seem to provide evidence in favour of my claim. Jou, Shanteau and Harris (1996), for instance, tested for "reciprocal answers". They asked subjects to motivate and justify their response to evaluative questions, and reported that these justifications are often inconsistent, showing an underlying lack of understanding of the scenarios.[21]

It is not my place here to discuss or criticize the methodology of such experimental surveys. I just note that the status of the experimental literature does seem to adequately answer the question of the source of empirical disagreement in fully-equalized cases. There are reasons to speculate, however, that more qualitative experimental settings, as well as investigating the prevalence and motivations of people resisting the task, may provide further support for the claim that most people do find these scenarios confusing, and that we should thus start questioning whether the whole strategy is cognitively and psychologically meaningful.

Trammell (1975, 1979), in his defence of the moral relevance of the killing/letting die distinction, makes some similar remarks. The case he discusses was introduced by Michael

[19] See, for instance, Barry, Lindauer and Øverland (2014), or Osman (2015).

[20] Greene et al. (2009), p. 184.

[21] Specifically, there is a significant number of "reciprocal answers". In an experimental setting involving the choice between two "fully-equalized" options, Jou, Shanteau and Harris (1996, p. 5) define "reciprocal answers" as justifications for one's choice which make explicit reference to the other option, by providing a rationale which could be equally applied for justifying *both* options. According to the authors, this result proves that agents are confused about the cases.

Tooley in 1972, and is identical in all respects to Smith/Jones. The author also argues that the lucky coincidence, and the fact that Jones cynically and satisfactorily observes his cousin drowning "have a ʻmasking' or ʻsledgehammer' effect, which makes it difficult to evaluate the significance of the distinction."[22] Nonetheless, "The fact that one cannot distinguish the taste of two wines when both are mixed with green persimmon juice, does not imply that there is no distinction between the wines." Similarly to my argument, Trammell claims that comparing fully-equalized cases can be extremely misleading, as it artificially pulls apart aspects of behaviours which, although distinguishable in principle, usually come together in our "practical moral life".[23]

In summary, it is reasonable to think that disagreement over fully-equalized cases depends, at least to some extent, on the fact that these scenarios are particularly difficult, unfamiliar and "unnatural". As a result, our moral intuitions over these cases do not seem particularly reliable as a test for the relevance of the doing/allowing distinction.

3.4 Conclusion

In this chapter I have argued that, while comparing fully-equalized cases seems, in principle, a useful tool for testing the positive and the negative theses regarding the relevance of the doing/allowing distinction, the most influential examples of this strategy do not live up to this promise. In the active/passive euthanasia and in trolley examples, the comparison of allegedly equalized cases appears to generate further disagreement over the correct classification of conducts, which cannot be easily separated from the issue of the moral relevance of the distinction. In this sense, this case seems to challenge the assumption that we can separate this latter question from the descriptive task of drawing a line between doing and allowing.

Furthermore, trolley cases amount to particularly difficult moral dilemmas, which are arguably too extreme and far from ordinary to count as reliable instances of moral intuitions. The Smith/Jones case, which seems to be adequately controlled, may also appear confusing and artificial.

The result of this survey leaves us with with the problem of explaining the pervasive lack of consensus. First, lack of consensus does not seem to be evidence in favour of either the positive or the negative thesis. Moreover, my suggestion is that the underlying strategy of fully-equalized cases, which assumes that we can actually disentangle this distinction from all other

[22] Trammell (1975).

[23] Ibidem, p.132.

morally relevant features of the context, might ultimately prove unfeasible.

4. Moral Framing

This chapter continues examining the issue of whether the doing/allowing distinction is morally relevant by considering the import of framing effects. First, I briefly describe what we refer to, in psychology and behavioural decision theory, as framing effects (4.1). In particular, I show how they seem to affect our doing/allowing classifications of apparently equivalent conducts. These results, arguably, raise some problematic questions for the reliability of our moral intuitions in general, and, more specifically, seem to challenge the positive thesis, that is, the claim that doing and allowing characterisations amount to morally significant features.

I appeal to an influential example in the moral literature, the "Asian flu" case, in order to examine in detail how empirical evidence regarding the persistence of framing effects impacts the positive and negative theses (4.2). On the one hand, it could be argued, following deflationist proposals such as Tamara Horowitz's and Walter Sinnott-Armstrong's, that framing effects prove that the doing/allowing distinction, under closer scrutiny, collapses into reasoning biases, such as the loss/no gain effect or the endowment effect. This interpretation thus seems to negate the positive thesis (4.3). A second position, upheld, among others, by Frances Kamm, claims instead that framing effects do not really affect the doing/allowing distinction once we examine these cases more thoroughly and control for cognitive biases. This conclusion is, therefore, compatible with the positive thesis, as it suggests that there could be a "correct" and morally relevant characterisation of doing and allowing that agents might fail to track in some specific situations (4.4).

In 4.5 I argue for third hypothesis: *contra* Horowitz and Sinnott-Armstrong, the doing/allowing distinction does not collapse into reasoning biases or behavioural effects described in the literature; *contra* Kamm, however, the doing/allowing distinction may still be subject to framing. This hypothesis, while not ruling out the positive thesis completely, seems to further challenge the idea that we can meaningfully disentangle the doing/allowing distinction from other descriptive and normative features of the context, and further suggests that doing/allowing classifications might ultimately be ambiguous.

4.1 Framing effects

A "framing effect" is generally said to occur when two descriptions of apparently equivalent decision problems induce systematically different responses and decisions.[1] This widespread phenomenon in choice contexts has been widely investigated in behavioural psychology and decision theory, following the seminal 1979 paper by Daniel Kahneman and Amos Tversky, and has been gathering strong empirical support. "Framing effects" can be caused by a variety of reasoning biases. Levin, Schneider and Gaeth (1998) report an exhaustive list of framing effects documented in medical decisions, bargaining settings, responses to moral dilemmas or in courts. In particular, Levin et al. suggest that framing effects can be classified according to *what* is being framed, and they distinguish between *attribute framing*, *risky choice framing* and *goal framing*. While I do not necessarily agree that these three categories are mutually exclusive or jointly exhaustive, they are arguably a useful tool for illustrating the most influential and investigated cases of framing effects.

Attribute framing refers to cases where an attribute or property of an object or an event is described in terms of either its positive qualities or its negative qualities, yet it can be easily inferred that the overall qualities are in each case the same. Empirical findings report that objects or events that are positively described are usually evaluated more favourably, or chosen over the negatively described objects. A typical example of this phenomenon is the preference for a product described as "25% lean" over a product described as "75% fat".

Risky choice framing refers to cases of describing options in terms of probabilities of achieving gains or losses. The different options are usually equivalent with respect to expected monetary value, but one amounts to a sure outcome while the other is a risky prospect/involves a gamble. The sure outcome and the gamble are "both described either in terms of gain outcomes and probabilities or else in terms of equivalent loss outcomes and probabilities".[2] Typically, agents tend to value more favourably or choose the sure outcome, and to have different risk attitudes with respect to outcomes depending on whether they are described in terms of probable gains or probable losses. These effects have been extensively examined and attributed to different features of human cognition, such as risk aversion, uncertainty aversion, status quo bias, loss aversion, endowment effect and so on. The main example I will appeal to in this chapter, the "Asian flu" case, is usually thought of as falling into this category.[3]

Finally, in *goal framing*, an activity is either described in terms of either the advantages of undertaking it or the disadvantages of not undertaking it, and agents are reported to prefer engaging when the latter are emphasized. This effect has been empirically observed in decision making about insurance policies, or in eliciting consensus on the criminalisation of certain

[1] Sher and McKenzie (2008), p. 1.

[2] Ibidem, p. 2.

[3] I do not necessarily stand by this characterisation as, arguably, the "Asian flu" case shares some relevant properties with all of these three classes of framing effects.

health or safety measures, such as wearing seatbelts.

Framing effects, as this brief summary shows, seem to amount to prevalent phenomena in many choice situations. As Shafir and LeBoeuf (2002) argue, framing effects are also conceived, in the standard narrative, as threatening and challenging to the traditional “rational actor model” and, in general, to the adequacy, reliability and rationality of human cognitive processes. It is not my place here to engage in what Shafir and LeBoeuf define as the “rationality debate”, but I simply note that, as Kahneman and Tversky observe, these kinds of preference patterns and decision behaviours seemingly violate the tenets of classical expected utility theory. Specifically, Kahneman and Tversky argue that framing effects are particularly problematic for the normative condition of *description invariance*, which requires that the same decision problem, in terms of expected utility, must be evaluated in the same way by any rational agent. Without dealing with further technical details, description invariance seems to have an intuitive appeal; arguably, there is something wrong with preferences being subject to framing effects, something that could lead us to question the adequacy of our reasoning processes, and to wonder whether they are reliable or rather in need of correction.

4.1.1 Framing doings and allowings

Let us now narrow the focus to the doing/allowing distinction. Cases of framing in this respect are the ones that involve moral evaluations of options, and where these moral evaluations seem to be affected by how a given conduct is described and presented to the agent. Specifically, I examine here the issue of whether framing effects can induce a different classification of the same action as an instance of doing rather than allowing, and different moral judgements of actions so perceived. If this were the case, we would then need to discuss the bearing of this phenomenon on the moral relevance of the doing/allowing distinction. Indeed, in the present section, I discuss how framing effects would bear on the positive *versus* the negative thesis, according to Horgan and Timmon's model of moral normativity and Sinnott-Armstrong's “master argument”. I also note, however, that supporters of the positive thesis could still argue that framing effects simply point to flaws in our moral reasoning.

Cases where our attributions of doing and allowing seem to be dependent on how the context is framed have been discussed by advocates of the negative thesis, such as Walter Sinnott-Armstrong and Shelly Kagan. Even without referring to framing effects in particular, Kagan (1989, p. 101) makes exactly this point when he asks the reader to compare the two following examples:

> My rival is going to win a prize, which would be automatically awarded to me if she were to die. Incidentally, my rival lies in a comatose state in the hospital, kept alive by a life-sustaining machine. I sneak into the hospital and unplug the machine, bringing about the death of my rival so as to win the prize.
>
> A doctor, after consulting with the parents and following the standard legal procedures, unplugs the machine that is keeping a comatose young boy alive.

In both cases, Kagan argues, the conduct is exactly the same, that is, unplugging a life-sustaining machine; the outcome is also the same, the death of the patient, and the action is arguably performed so as to bring about the patient's death. Nonetheless, the different context our respective conducts is put into induces a different description of my behaviour and the doctor's: while the former action is usually described as "killing" my rival, the latter is perceived as "letting the boy die", and the latter is evaluated more favourably than the former. Another common example is the comparison between refusing to give money to a starving beggar and starving one's baby to death: again, while the two actions are seemingly extensionally equivalent, most people tend to agree on the characterisation within the former frame as allowing a harm, but change their classification in the second frame, describing the conduct as "harming one's baby" (doing harm).

The significance of these examples of alleged framing effects, however, has been criticized by supporters of the positive thesis. Arguably, in Kagan's example, despite the conduct "unplugging the machine" being identical in both cases and amounting to the same way of bringing about a harm, the two scenarios seem to describe two different moral problems. Rather than a mere change of "frame", indeed, we are confronted here with two vastly different narratives and scenarios, and it is doubtful whether they could be effectively conceived of as "equivalent". The background norms in this pair of cases are obviously different. The selected doing/allowing account I discuss in Chapter 2 thus equips us with adequate tools to explain why these two conducts amount to "doing" and "allowing" respectively. In this section, therefore, I sideline these cases where norms play an obvious role, and I focus on less "moralised" examples, which appear to be ultimately instances where the doing/allowing distinction is subject to framing.[4]

The more straightforward cases refer to experimental settings where agents are faced with the same decision problem – in terms of outcomes and actions – which is simply described using different words, or where the order of sentences is changed. Petrinovich and O'Neill (1995), for

[4] In Chapters 5 and 6 of this thesis, I will argue that norms do, in fact, play a role, even in less "obviously" moralised cases. In the present chapter, however, I strive to avoid at least the most straightforward examples in this respect.

instance, analysed people's responses to a trolley case where they are asked to identify with the bystander who could either let the trolley follow its track and run over five people or throw the switch so that the trolley goes to a side track, running over one person (that is, the problem described in Bystander). Respondents were asked to evaluate the two conducts on a 6-point scale from “strongly agree” to “strongly disagree”. In one group, the options in the trolley case were described using the word “kill” – throw the switch and kill one person or let the trolley stay on track and kill five – while the second group worked with questionnaires where the options were described as “saving” – turn the trolley and save five persons or do nothing and save one. Empirical surveys reported that agents were “likely to agree more strongly with almost any statement worded as Save than one worded as Kill”.[5] Specifically, people were more likely to agree, and agreed more strongly, that throwing the switch was permissible, and morally preferable, when this conduct was characterised as Saving. While people still judged that it was permissible to Switch, it seems that they felt more comfortable with and sure of their decisions when the wording was stated in terms of “allowing”.

Sinnott-Armstrong (2008) surveys other similar experimental settings; specifically, he argues that empirical data seems to show that moral judgements can also depend on other framing effects besides wording, such as the order in which the examples are presented to the reader.[6] In another questionnaire by Petrinovich and O'Neill, for instance, the Bystander case was put in a list together with two other examples, and the agents were again asked to evaluate these cases. The other examples are the following “scan case”:

> The only way to save five dying persons is to scan the brain of a healthy individual, thus killing that innocent person.

And the “transplant case”:

> The only way to save five people is to transplant organs from a healthy person, thus killing that innocent person.

All of the options were always described using the “saving” wording this time (“save” five people by killing one), but were presented to different groups of participants in a different order. The authors reported that answers did change dramatically, and, specifically, for Switch in the trolley case, “people more strongly approved of action (throwing the switch) when it appeared last in the sequence than when it appeared first”.[7] This conclusion seems to indicate that the comparison with “similar” but morally worse scenarios affects how people judge the same action “throwing the switch”. In particular, the transplant case and the scan case appear to be

[5] Petrinovich and O'Neill (1995), p. 149.

[6] Ibidem, p. 152.

[7] Ibidem, p. 155.

more clear-cut instances of doing harm, and thus are more likely to be considered morally impermissible, even compared with the possibility of saving lives. Switch tends to be seen in a less favourable “doing” light when evaluated independently of (before) the obviously worse doing cases, and in a more favourable “allowing” light when evaluated after the obviously worse doing cases.[8]

4.1.2 Framing and the reliability of moral intuitions

The findings above seem to point to the fact that, when asked to morally evaluate conducts that are apparently equivalent, both our characterisations of behaviours as doings rather than allowings and/or our judgements about the relative blameworthiness of these cases may be affected by framing effects, such as the use of specific words or the order in which cases are presented. But are framing effects problematic? Intuitively, it is quite discomforting to realize that our moral judgements depend on seemingly morally irrelevant features, such as the mere use of the word “killing” rather than “saving” or of equivalent, but differently phrased, probabilistic information. As Kahneman and Tversky explain for non-moral cases, some normative feature of our reasoning appears to be violated by the shift in value judgements framing effects bring about; similarly, for moral cases, this violation seems to point to some inconsistency in our moral reasoning processes.

Horgan and Timmons (2009), specifically, argue that framing effects threaten what they define as “moral normativity”, that is,

> “the fact that the principle should connect descriptive features to moral-normative features in such a way that for any potential circumstance C, all true moral-normative statements about C (e.g., statements about what would be morally just in C, what would be morally obligatory in C, what would be morally wrong in C, etc.) are logically entailed by a conjunction of (i) a sufficiently detailed characterization of C in descriptive, non-normative, terms and (ii) the sought-for principles”.[9]

This requirement, which amounts to a specific kind of internal consistency, would seem to be openly violated when framing effects are at work. Two extensionally equivalent decision problems can, indeed, be described as the same circumstance C, thus fixing (i). We then add to C all the morally relevant principles, thus fixing (ii). However, the result of this process is two

[8] This is not surprising in light of my discussion in Chapter 3: whether Switch in Bystander amount to doing or allowing seems at least controversial.

[9] Horgan and Timmons (2009), p. 26.

different moral evaluations of C. This result is, clearly, problematic insofar as these moral/normative statements are not logically entailed by the conjunction of the descriptive features of the moral dilemma (i) and the exhaustive "set" of moral principles we employ (ii). Horgan and Timmons thus conclude that framing effects show some kind of flaw or inconsistency in our moral reasoning. The subsequent question for Horgan and Timmons concerns what kind of cognitive processes are at play when we make moral judgements, given that the straightforward model provided above fails to capture the connection between descriptive features, fixed moral principles and resulting moral judgements.

In this chapter, I narrowly focus on whether and how framing effects could impact the negative and the positive theses regarding the moral significance of doing and allowing. At first sight, the pervasiveness of framing effects in our doing and allowing characterisations could be interpreted as bad news for the positive thesis.[10] After all, if the way we employ this distinction is unstable, and relies heavily on seemingly "irrelevant features", which can be manipulated through framing, the claim that this dichotomy holds some intrinsic moral relevance appears to be seriously undermined, as per Horgan and Timmon's analysis. The positive thesis regarding the doing/allowing distinction, in fact, at least partially relies on intuitions about examples, and these intuitions are assumed to be reliable. Specifically, the positive thesis is grounded in the strong intuitive appeal of the principle "doing is worse than allowing", and in the fact that we persistently employ it in real life. Framing effects, as Sinnott-Armstrong (2005) puts it, nonetheless appear to question whether these moral intuitions are justifiable at all.

Sinnott-Armstrong[11] provides further details as to why framing effects undermine evidence from intuitions about both the doing/allowing distinction and its moral relevance. The general process we use to justify a certain belief, Sinnott-Armstrong claims, is in fact a kind of inferential confirmation: in other words, we show the circumstances "whose denial undermines it", and then show that the belief was not formed under such circumstances. Suppose, for example, that I believe my mother will arrive at Heathrow at 5pm, because she told me so yesterday. In this case, the justification for my belief appeals to the general reliability of the process of basing my credence on this empirical evidence. Imagine now that I realise that yesterday I took a potent hallucinogenic drug. These circumstances, of course, undermine my belief that my mother will arrive at Heathrow at 5pm, because the process of forming beliefs under the effect of hallucinogenic drugs is not generally reliable; the evidence that I took the drug, therefore, stands

[10] Of course, this investigation is meaningful only insofar as there are real examples where our doing/allowing characterisations are subject to framing. The trolley example above seems to amount to a convincing example of frame-dependency. The most straightforward example in this respect, however, is arguably the "Asian flu" case, which I discuss later in this chapter.

[11] I refer here to Sinnott-Armstrong's discussion outlined in "Framing Moral Intuitions", pp. 48–58.

as a *defeater* for my belief. The fact that I took the drug, in other words, denies the general circumstances under which my belief would be justified. Certainly, I could still provide justification for my belief by appealing to other sources: I could call my father, asking him whether my mother will arrive at Heathrow, or I could go to the airport at 5pm and check whether my mother is there. The evidence that my beliefs are formed using unreliable processes, however, requires a separate justification for those beliefs. To sum up, Sinnott-Armstrong claims that an adequate justification for my belief amounts to showing that some circumstances that would undermine it do not occur. If, on the other hand, the agent is aware that these defeating circumstances are in place, the belief needs to be supported by additional justifications.

Applying this framework to our moral intuitions, Sinnott-Armstrong sets out the following "master argument": "if our moral intuitions are formed under circumstances which are not reliable (...), then our moral intuitions are not justified without further confirmation".[12] In the case of the doing/allowing distinction, the persistence of framing effects seems to show that our intuitive classifications and judgements are formed under unreliable circumstances, and that the general process we use to perform such judgements is thus itself unreliable. In the trolley case, for instance, evidence that changing the wording from "killing" to "saving" induces a shift in people's responses is a *defeater* for the moral intuitions serving as evidence one may have about either of the cases. This may seem to support the negative thesis – that doing/allowing is not a morally relevant distinction, if a proper distinction at all.

This sceptical position, though reasonable, can still be challenged by supporters of the positive thesis. While framing effects show that our moral judgements and intuitions are not always reliable and consistent, it could be argued that there is still a fact of the matter whether a conduct is an instance of doing rather than allowing, and that this characterisation of conduct is morally significant *per se*. Framing effects could just be evidence of the fact that our first-hand intuitions and moral evaluations are not (always) to be trusted, and that we should be more careful when appealing to them in our theoretical efforts to make sense of the doing/allowing distinction. Roughly, the fact that in some cases our moral reasoning can be swayed by framing does not prove that we could not build an adequate account of the doing/allowing distinction, relying on the "correct" kind of intuitions. Going back to Sinnott-Armstrong's master argument, supporters of the positive thesis could still argue that our intuition that "doing harm is worse than allowing it" could be backed up by other external justifications, such as some particularly trustworthy and strong intuitions, analogous to calling my dad in the Heathrow example.

[12] Ibidem, p. 52.

In the following sections, I examine the classic "Asian flu" case. As a convincing example of framing effects, I use it to outline different hypotheses and explanations of the bearing of framing on the moral significance of the doing/allowing distinction. In 4.2, I show how Kahneman and Tversky's experimental results are commonly described in terms of the so-called "loss/no gain effect". I then examine how these results can be explained by both deflationist (4.3) and non-deflationist (4.4) hypotheses. The former accounts for evidence of framing effects in a way which is consistent with the negative thesis, while the latter does so in a way that is consistent with the positive thesis. I conclude that a hybrid hypothesis remains open (4.5).

4.2 The "Asian flu" case

In their 1983 paper "Choices, Values and Frames", Daniel Kahneman and Amos Tversky introduce the concept of a decision frame and outline the tenets of Prospect Theory, which they regard as a model of how agents actually choose. By way of supporting their proposal, they report and analyse different empirical results, among which is the famous "Asian flu" case. This experimental setting divides the subjects into two groups; the first is faced with the following dilemma:

> Your city is threatened by an "Asian flu" that is expected to kill 600 people, and you have to make a choice between these two alternative vaccination programs:
>
> - If Program A is adopted, 200 out of the 600 people will be saved.
> - If Program B is adopted, there is a 2/3 probability that no-one will be saved and 1/3 probability that all 600 people will be saved.
>
> Which program would you choose?

The second group was faced with the very same scenario, but the choice was instead between C and D:

> - If Program C is adopted, 400 out of the 600 people will die.
> - If Program D is adopted, there is 1/3 probability that no one will die and a 2/3 probability that 600 people will die.

A and C, like B and D, are clearly extensionally equivalent with respect to lives saved, and describe the same vaccination program: "200 people will be saved and 400 will die" (A and C) and "there is 1/3 probability that 600 people will be saved and no one will die and a 2/3 probability that no one will be saved and 600 people will die" (B and D). Therefore, we could reasonably expect that the percentage of people opting for A and C (or for B and D) would be similar in the first and second groups. Nonetheless, experimental findings showed that 72% of subjects in the first group chose Program A but, in the second group, 78% of subjects chose Program D.

Kahneman and Tversky use the Asian flu case, together with five other experimental settings, as representative examples of how *decision frames* affect agents' behaviours. By a decision frame, the authors mean "the decision-maker's conception of the acts, outcomes, and contingencies associated with a particular choice"[13]. As "the frame that a decision-maker adopts is controlled partly by the formulation of the problem and partly by the norms, habits, and personal characteristics of the decision-maker"[14], the same decision problem will often have several decision frames, even for the same decision maker (thus holding fixed the personal characteristics and changing only the formulation). In particular, in the Asian flu case, they argue that the two different decision frames do not involve different factual descriptions of the world, but rather assume a different reference point as the *baseline*.

By way of elaboration: according to standard expected utility theory, the change in a decision frame, or, more specifically, a change in the reference point, should be irrelevant to the choice of a course of action. Kahneman and Tversky, however, point out that agents do not seem to choose this way: specifically, they do change their behaviour depending on the decision frame, as the "Asian flu" case clearly illustrates. This violates consistency norms rational decisions should supposedly conform to, as the two frames are, with respect to standard decision theory, two descriptions of situations that are identical in all relevant aspects.

Let's now see in detail how reframing supposedly explains preference reversal in the "Asian flu" case. Kahneman and Tversky argue that 1) the reference point matters for choice behaviour and 2) people are generally more risk seeking when it comes to avoiding sure losses from a given baseline, and more risk averse when it comes to pursuing gains from a given reference point.[15] In Kahneman and Tversky (1983) and in Kahneman, Knetsch and Thaler (1990), the difference in risk attitudes summarised in 2) is further examined and associated with a psychological mechanism known as *endowment effect*. With the support of further empirical evidence, the authors observe that agents prefer to avoid losses rather than to acquire equivalent gains, as they seem to value an object (or an amount of a given currency) more if they already "own" it or feel

[13] Kahneman and Tversky (1983), p. 455.

[14] Ibidem.

[15] Note that there is a further crucial element of Prospect Theory – an extra risk parameter that modifies the probability contribution to the evaluation of an option. This nonetheless amounts to a separable component of the theory that is not the focus here.

somehow attached or entitled to it; this, arguably, makes it worse for them to lose it with respect to the enjoyment they would experience in gaining the same object (or amount of currency).[16] As a result, people are less prone to take the risk of improving from the baseline and more prone to take the same risk to avoid getting worse compared to the baseline.

Cases of endowment effect are ubiquitous and already familiar from bargaining settings, with respect to agents' willingness to pay *versus* willingness to sell. One famous example of the psychological pull of the endowment effect, for instance, is the so-called Knee example.[17] Kahneman and Tversky polled subjects to ask what amount of money they would demand a) *in compensation* for not getting a lost knee back and b) *in exchange* for losing a knee. The results show that agents demand more money ex ante (b), i.e., when they are faced with the possibility of losing their knee, than they demand ex post (a), i.e., when they are told they don't have a knee and could get it back. Apparently, this supports the conclusion that people value the knee they already have more than the knee they would get back, even assuming that the new knee would be equal to the old one in all respects.

Let's now take a closer look at how the endowment effect would work as an explanation of the "Asian flu" case.[18] In the first decision problem (the choice between A and B), the use of the phrasing "saving" identifies the 200 lives as a gain, thus seemingly setting the reference point at "all 600 people die". With respect to the baseline "everyone dies", choosing program A would amount to a sure gain from the reference point. Plan B, on the other hand, characterises a "bet", as it involves evaluating a risky prospect. Specifically, with respect to the baseline "everyone dies", Plan B could either deliver a bigger gain (all 600 people saved) or simply make no progress at all from the baseline (all 600 die). When it comes to gains, Kahneman and Tversky observe, decision makers tend to be risk averse, and, given the same expected lives saved in A and B, most opt for Plan A, which guarantees a sure gain. In the second decision problem, the different framing of the decision triggers a different evaluation of the vaccination plans. Plan C, indeed, apparently presents the option of 400 people dying as a loss, as it uses the phrasing "die"; this description thus sets the baseline at "all 600 people live". With respect to this reference point, Plan C therefore involves a sure loss. Plan D, again, amounts to a bet, where either losses with respect to the baseline are completely avoided (no one dies) or a bigger loss could occur (all 600 people die). While C and D are expected-lives-saved equivalent, decision makers mostly opt for D, being risk-loving with respect to losses.

[16] Kahneman, Knetsch and Thaler (1990), p. 195.

[17] Frances Kamm (2007), pp. 472–73, also compares the "Asian flu" case with the Knee case, as we will see later.

[18] Kahneman and Tversky do not discuss this interpretation in detail; what follows is thus my reconstruction of the case.

In conclusion, according to Kahneman and Tversky, different framings select different reference points as the relevant baseline, namely "everyone dies" *vs* "everyone lives", and this, in turn, induces a different perception of the options as gains rather than losses. Because of the endowment effect, agents would then tend to value the same numbers of lives more when they feel they already "own" them (or they feel they are already secured); therefore, people are supposedly risk seeking when it comes to avoiding losing lives that are framed as losses with respect to the reference point "everyone lives", and risk averse when it comes to saving lives that are framed as gains from the reference point "everyone dies". Consistently, they tend to choose the course of action that involves a chance to completely avoid any loss (plan D over plan C), but are not as eager to take the same risk to save more lives (plan A over plan B).

Kahneman and Tversky's explanation of the peculiar behaviour observed in the Asian flu case, which has become the standard narrative in choice theory circles, thus amounts to a combination of different attitudes and features of human reasoning: baseline sensitivity, endowment effect and the related "loss/no gain" effect. By triggering and manipulating these features, shifts in framing cause the preference reversal. Kahneman and Tversky conclude that these preference reversals are, in fact, irrational, but argue that Prospect Theory can descriptively account for them.

4.3 The deflationist hypothesis

In their 1983 paper, Kahneman and Tversky do not talk about doing and allowing. Subsequent discussions of the Asian flu case in the moral and behavioural literature, however, have often related these experimental findings to the doing/allowing distinction. Specifically, this case is often referred to as a straightforward instance of framing that affects doing/allowing classifications. The Asian flu case, in fact, describes a situation where a harm (death of innocent people) causally depends on the way the agent chooses between options, whether by doing the harm or allowing the harm to occur. It is reasonable, therefore, to think that agents could be discriminating the available options in terms of doing or allowing, and appealing to the idea that "doing is worse than allowing" when making evaluations. Preference reversal, in turn, would indicate that agents make different doing/allowing characterisations of apparently extensionally equivalent behaviours.

Note that different hypotheses can be put forward to account for the fact that, in the Asian flu case, doing/allowing classifications are seemingly frame-dependent. In the present section, I analyse what I call *deflationist* hypotheses, i.e., positions that, to some extent, argue that frame-dependency undermines the moral relevance of the doing/allowing distinction. In section 4.4, I survey Kamm's *non-deflationist* hypothesis, which claims that framing effects do not threaten the normative significance of the distinction. In this sense, these hypotheses amount to different positions on the import of evidence of framing on the negative/positive theses. Deflationist hypotheses, specifically, explain framing in a way that supports the negative thesis, while non-deflationist ones provide an explanation of framing effects in a way that is consistent with the positive thesis. In 4.5, I turn to my interpretation of Kahneman and Tversky's results, which is that the doing/allowing distinction does not collapse into the loss/no gain effect, but that doing/allowing classifications are sensitive to framing, and thus allow for ambiguity.

In the deflationist camp, Tamara Horowitz and Walter Sinnott-Armstrong argue that Kahneman and Tversky's explanation of preference reversal proves that intuitions apparently concerning doing/allowing turn rather into reasoning biases such as the loss/no gain effect and the endowment effect. Specifically, the way in which we use doing and allowing classifications may be expressed in terms of the cognitive attitudes we exhibit in dealing with gains *versus* losses.

Horowitz, in particular, builds upon Kahneman and Tversky's explanation: whether an outcome is perceived as a loss or a no-gain depends on the perceived baseline, which itself depends on framing effects (wording); because of the endowment effect, agents place greater value on lives already secured; due to the loss/no gain effect, they are more risk seeking when it comes to avoiding losses and more risk averse when it comes to pursuing gains. At this point, Horowitz argues, a further element complicates the picture drawn by Kahneman and Tversky: in a morally sensitive scenario, such as the Asian flu case, agents classify losses from a given baseline, which causally depends on their behaviour, as doing harm, while they classify no-gains from a different baseline, which they otherwise brought about in the same way, as allowing harm. The different risk attitudes analysed by prospect theory then justify why agents tend to evaluate more favourably the same option when it is framed as a no-gain rather than a loss, which in turn explains, as a result or byproduct of these risk attitudes, why agents consider doing harm to be morally worse than allowing harm to occur.

This account explains the doing/allowing distinction in purely non-moral terms and, what is more, as the effect of psychological attitudes, idiosyncrasies and reasoning biases. Therefore, the positive thesis would be undermined, as we have a better explanation of the intuition that "doing is worse than allowing" (namely, that the "worseness" of doing is a by-product of risk attitudes). Furthermore, the doing/allowing distinction would build on our flawed moral reasoning skills.

There are other interpretations of the Asian flu case in the literature that similarly seem to show that the doing/allowing distinction can be explained as a reasoning bias. Christian List and Nathalie Gold (2004), for instance, examine the Asian flu case in a "dynamic reasoning" model, which shows preference reversal to be the result of an effect which they call path-dependence. While List and Gold do not actually mention doing *versus* allowing, their model may provide yet another explanation for the fact that agents might evaluate the vaccination plans differently. In fact, one plan could be judged as morally worse than another simply because of the different order (or, to use their terminology, decision path) in which background propositions are arranged.[19]

This account, even if not explicitly "deflationist" like Horowitz's, may be used to argue in the direction that doing/allowing classifications allow for ambiguity, as both characterisations are tenable depending on which decision path is elicited. Arguably, if the selection of one path over another is induced by a change of words, which seems morally irrelevant, we could conclude that these classifications are also not morally significant, or at least that our intuitions regarding doing/allowing are not to be trusted.

4.4 Non-deflationist hypothesis: Kamm's proposal

In this section, I turn to Kamm's response to the deflationist challenge. As discussed above, deflationist approaches seem to amount to bad news for the positive thesis, as they seek to explain the doing/allowing distinction by referring to cognitive or behavioural features that do not have any obvious moral relevance, and may even be irrational biases. Among others, Frances Kamm has famously argued against this conclusion, challenging Sinnott-Armstrong's and Horowitz's positions. In what follows, I present in what I take to be the most favourable light Kamm's contribution to the debate as outlined in her 1998 paper "Moral Intuitions, Cognitive Psychology, and the Harming-versus-Not-Aiding Distinction" and in Chapter 14 of "Intricate Ethics" (2007). This is not the most faithful representation of all the twists and turns in Kamm's analysis (although I will later note some of the more detailed points that Kamm makes that are difficult to square) but I think it is a charitable representation of Kamm's view.

[19] List and Gold (2005, p. 6) define background propositions as the decisional context of the agent, and, more precisely, as all the elements the agent considers when forming a judgement about a "target proposition" of the form "*x* is preferable to *y*". These background propositions may be either factual or normative propositions, which represent moral principles such as, for instance, "do not kill on purpose". Background propositions, in short, are the propositions an agent refers to when motivating her decision on or evaluation of the target proposition. If agents do accept inconsistent background propositions, they could either accept or reject a target proposition, depending on which subset of background propositions the frame makes focal.

In short, Kamm argues that the impact of framing effects can be properly downplayed, and thus that our intuitions can serve as an adequate account of the doing/allowing distinction. Specifically, I take her argument to consist of two main claims: (a) the doing/allowing distinction is different from the loss/no gain distinction, and thus the former cannot be reduced to the latter; (b) the doing/allowing distinction, when correctly characterised, is immune to framing effects.

Kamm begins her discussion by exposing a "hidden agenda" behind Kahneman and Tversky's loss/no gain interpretation of cases such as the Asian flu.[20] This hidden agenda, as analysed by Kamm, has the following structure: the doing/allowing distinction is largely supported by people's intuitions about cases; in some instances, these intuitions seem to exhibit preference reversal, which seems to be convincingly explained by the loss/no gain effect; the loss/no gain effect has no obvious moral relevance, as it depends on apparently morally irrelevant features, such as the choice of framing; this eventually undermines the moral relevance of the doing/allowing distinction. Note that, if this argument stands, we will have a compelling case in favour of the negative thesis.

Kamm's first move in rejecting this argument is that our moral intuitions about cases cannot be adequately explained by the loss/no gain effect. As such, the doing/allowing distinction does not collapse into the loss/no gain distinction (a), as the latter cannot fully account for our discrimination between doing *versus* allowing cases. As a preliminary observation, Kamm notices that while losses and gains are descriptions referring to end-states, doing and allowing have to do with the way in which an agent brings about these end-states. As such, "doing" and "allowing" identify the connection between the agent's behaviour and the outcome. Most examples used to show that the loss/no gain and the doing/allowing distinctions are equivalent, Kamm argues, overlook this aspect. In the "Asian flu" case, for instance, Kahneman and Tversky's and Horowitz's analyses are misleading insofar as they simply refer to the vaccination plans in terms of gains and losses.

To make this point, Kamm distinguishes between different ways in which losses and gains can occur, by referring, for instance, to *unprevented losses*[21] and *denied gains*. I do not follow Kamm's discussion in detail here, as it is not always clear into which category the different vaccination plans would fall. In short, Kamm argues that Horowitz's analysis of the doing/allowing distinction is too simplistic. To be fair, in most scenarios doing harm is associated with cases we perceive as suffering losses and allowing harm with cases we perceive

[20] Note that Kahneman and Tversky do not suggest that the loss/no gain distinction undermines the doing/allowing distinction.

[21] Kamm defines unprevented losses as "losses which are happening independently of our intervention and which we fail to prevent" (1998, p. 475).

as not achieving a gain. This is to say, the loss/no gain and the doing/allowing distinctions often match up. Nonetheless, this is not always the case, and we can think of ways to disentangle the two distinctions. By way of illustration:[22]

	Doing Harm	Allowing Harm
Loss	I roll your apple down the hill	I don't stop your apple from rolling away
No-gain	I divert an apple which was rolling towards you	There is an apple rolling past you, and I don't divert it towards you

Table 4.1

In other words, it is possible for a victim to suffer a loss as a result of someone's allowing harm (as per the top right entry of Table 4.1). I take this to be what Kamm means to capture by "unprevented losses". Also, it is possible for a victim not to achieve a gain as a result of someone's doing behaviour (as per the bottom left entry in Table 4.1); Kamm seems to call these cases "denied gains".[23]

Now that it is clear that the two distinctions do not collapse into one another, but often match up in ordinary cases, Kamm turns to explain the preference reversal, that is, why "subjects think it is worse if two hundred people lose life than if they do not gain it and are more averse to a policy in which people lose their lives than in which the same number are not saved."[24] According to Kamm, all the options involved in the Asian flu case amount to "allowing harm": what is actually bringing about the deaths is the flu, which is a natural event we could not possibly prevent. Different baselines induce a different perception of the same outcomes as gains or losses, and it is the loss/no gain effect that ultimately causes preference reversal. All cases are, nonetheless, still instances of allowing harm, and the doing/allowing distinction, if properly analysed, is immune to framing effects (b).

To illustrate this point, I reconstruct Kamm's analysis of the Asian flu case.[25] In this example, Kamm argues, confusion might arise because often the most natural and intuitive way to think about doing and allowing is to refer to human intervention, which in turn can be defined in comparison with a baseline, which sets what is going to happen absent the agent. In the Asian flu case, Programs A and C individuate different baselines, "everyone dies" *vs* "everyone lives",

[22] Thanks to John Cusbert for this example to illustrate the taxonomy.

[23] Note that, in a different passage, Kamm instead classifies "denied gains" as a case of allowing harm.

[24] Kamm (1998), p. 466.

[25] Also note that Kamm's analysis of this example is not a unitary explanation. In this paragraph I thus put together the remarks that Kamm makes throughout her paper.

and therefore the "absent the agent" situation is set differently in those cases. As such, Kamm argues, what we perceive as intervention might also appear different: in Program A, it looks like *with* our intervention 200 people could be saved from sure death; in Program C, it looks like *with* our intervention, 400 people will die. This could generate the confusion that our intervention amounts to a different thing or contribution to the outcome in Programs A and C respectively. Upon careful scrutiny, however, we can properly recognise that what does the trick here is just the loss/no gain effect.

In the first decision problem agents are, in fact, confronted with a "near death state",[26] where "everyone dies". The resulting lives saved, with respect to this reference point, are thus perceived as gains, but of a specific kind that Kamm defines as "maintaining what one is close to los[ing]".[27] The 400 lives lost, in this scenario, would then be no-gains of the form "*not* maintaining what one is close to lose".[28] I take this case to be in the bottom right entry in Table 4.1. On the other hand, the second decision problem (C *vs* D) sets the baseline at "everyone lives". With respect to such a situation, the 400 lives lost are framed as losses, which Kamm defines as "losses which are happening independently of our intervention and which we fail to prevent", i.e., unprevented losses. It is this different description of no-gains as opposed to losses that, ultimately, causes the different evaluation of decision problems, and could generate some confusion about the correct doing/allowing description. Nonetheless, Kamm concludes, this is just the loss/no gain effect; doing/allowing classifications would not actually be affected by framing (both these types of no-gains and losses are in the "allowing" camp).

Moreover, upon reflection, Kamm argues that only one baseline is "correct": namely, that "everyone dies". When we correctly identify the baseline, we can completely "beat" framing effects. Nonetheless, on Kamm's account, different baselines do not change the classification of behaviours as doings *versus* allowings. To be sure, doing/allowing classification may *seem* frame-dependent: once we realise our error, however, we will see that doings can be distinguished from allowings in a reliable way by appealing to a single baseline that is not subject to change due to differences in wording, order, etc. The import of framing effects is thus downplayed to a masking or misleading impact on our initial classifications.

In conclusion, if Kamm is successful in refuting the "hidden agenda", framing effects would not threaten the positive thesis regarding the moral relevance of the doing/allowing distinction.

In the following section, I suggest that, while I agree with Kamm that the doing/allowing distinction does not collapse into the loss/no gain effect, we can still argue that doing/allowing

[26] Kamm (1998), p. 466.

[27] Ibidem, p. 475.

[28] In a different passage, however, Kamm seems to define A *vs* B as a case of denied gains.

classifications are frame-dependent, and two different doing/allowing descriptions might be legitimate. My different interpretation builds on what I think is a gap in Kamm's analysis above. If doing/allowing and loss/no-gain are indeed different, we still do not have a compelling explanation for why framing effects can also change the perception of how losses and gains are brought about by the agent, which would amount to the doing/allowing distinction.

4.5 The third path

The comparison between deflationist and non-deflationist proposals seems to show that framing effects do not provide a decisive test in favour of either the positive or the negative thesis. Both narratives can, in fact, arguably account for the empirical data on preference reversals. In this section, I suggest a “third path”, with respect to both deflationist and non-deflationist hypotheses. I argue that this third interpretation is as well consistent with the empirical evidence; also, it may advance discussion and provide some explanatory advantages. On the downside, this interpretation seems to further challenge the feasibility of the project of fully isolating the doing/allowing distinction from all other descriptive and moral features of the decision context.

Firstly, contra Horowitz, I agree with Kamm that preference reversal in the Asian flu case cannot be fully expressed as the result of the endowment effect and the loss/no gain effect. The Asian flu case is in some respects different from other examples surveyed by Kahneman and Tversky (1979 and 1983), such as decisions about lotteries or the classic Knee case. For instance, in the latter, which is crucial for introducing the endowment effect, the knee is an object, or, more accurately, a possession, which agents may attain or be deprived of without any further specification of how the loss was/is brought about. In the Knee case, the agent thus has to choose between different outcomes that are only affecting her. The Asian flu case, on the other hand, also asks the agent to put herself in the hypothetical position of selecting between vaccination programs that will affect other people. I argue that, for this reason, the issue of how outcomes are brought about becomes more relevant. The knee, or its absence, can thus be described as a loss or a gain from a given baseline. When considering the number of lives lost or saved in the “Asian flu” example, instead, I think agents also focus on what kind of impact they had by bringing about the outcome. This aspect, I argue, may not be completely captured in loss/no gain terms.

While Kamm's objection (a) to Horowitz's account thus appears to be convincing, it is nonetheless questionable whether she can prove decisively that the doing/allowing distinction is immune to framing effects (b). *Contra* her conclusion, I argue that, in the Asian flu case, our perception of vaccination plans as instances of doing harm rather than allowing harm ultimately causes the preference reversal.

First, Kamm's reconstruction of the Asian flu case does not successfully illustrate how changes in the baseline affect agents' relative evaluations of the vaccination plans. Recall that, according to Kamm, all options involved in the case would be instances of allowing harm. Agents' different judgements and choices would be motivated by psychological biases such as the loss/no gain effect. If the two dichotomies (loss/no gain and doing/allowing), however, amount to distinct things, it is not clear how, on Kamm's account, people could still classify the options differently in terms of doing/allowing: after all, individuals should only describe the options differently in terms of gains and losses, and not in terms of how these outcomes are brought about by the agent. In this respect, I argue that Kamm's proposal lacks explanatory power.[29]

To fill this gap, I thus provide an explanation of how different baselines may induce different descriptions of vaccination programs as doings or allowings. My interpretation shares the same structure as Kahneman and Tversky's, with respect to baseline sensitivity and risk attitudes, but I suggest that it is the principle "doing is worse than allowing", rather than the endowment effect, that does the explanatory work. Specifically, the two different framings of the same decision problem induce a different perception of the same option as an instance of "doing harm" rather than "doing good while allowing some harm as a side effect"; as agents are influenced by this classification, and regard doing harm as morally worse than allowing it, they are risk seeking when it comes to avoiding doing harm (by causing a certain number of deaths), and risk-averse when it comes to doing more good, allowing harm as a side effect (by not saving the same number of lives).

By way of elaboration, the explanation of preference reversal is as follows: for plans A and B, as already discussed, the appropriate baseline seems to amount to a situation where everyone dies; agents, as Kahneman and Tversky argue, tend to take such a baseline as the position they find

[29] To be clear, there are no empirical surveys that specifically show how people change their doing/allowing classifications in the Asian flu case. We could thus simply conclude that the doing/allowing distinction has nothing to do with the Asian flu case. Nonetheless, there is evidence that people think about these options in terms of killing/letting die, and justify their answers by appealing to such descriptions. For instance, Jou, Shanteau and Harris (1996, p. 5) report individuals' motivations for choosing a specific option in the Asian flu case. These motivations seem to explicitly refer to these different ways of describing a conduct: e.g. "I wouldn't be able to live with myself knowing that I let 400 people die" (for an individual who chose Plan B over A) or "I would rather take a chance to save all of them than *sending* [my italics] 400 of them to death" (for choosing D over C). I thus side here with Horowitz and Kamm that the Asian flu case does involve doing/allowing descriptions of the vaccination programs. If this is the case, we do need to provide an explanation of how people could perceive these seemingly equivalent conducts differently (in doing/allowing terms).

choosing plan A, therefore, arguably appears from the agent's perspective to actively save 200 people, and *cause* the fact that they will live. On the other hand, the fact that the other 400 people will die is taken to be a side effect of the good action we are performing: in other words, it sounds more as if we are allowing 400 people to die (certainly, a harm), while causing 200 people to live. Plan A, in short, may be interpreted as "doing good and allowing harm (for sure)". As for plan B, agents must take a gamble: they could do even more good, completely eliminating the chance of even allowing harm, or they can end up allowing more deaths to occur. While equivalent with respect to expected harm (lives lost), most agents stick to plan A, as they are risk-averse when it comes to doing good, while allowing harm to occur. This, I argue, happens because, in this specific choice problem, option A is framed as giving the agent the chance to do something good for sure, with respect to the reference point; the harm that will occur as a consequence of her choosing A is framed in a way that appeals to "allowing", and thus to an occurrence of harm that is not as wrong as "doing". On the other hand, B could offer the chance to do even more good, but at the risk of not doing any good at all. I suggest that I suggest that, when it comes to allowing harm, agents are risk-averse.

Turning to the choice between C and D, the implied baseline is now "everyone lives". Plan C, a vaccination plan described as "killing" 400 individuals, appears in that respect to certainly *cause* the death of 400 individuals, thus actively harming them. In other words, I argue that the terms used in the description of Plan C are easily associated with a way of bringing about harm that is characterised as "doing". D, again, describes a lottery where the agent could end up doing even more harm (600 people die), but also not harming anyone (600 people live). Plan D, in this sense, avoids "doing harm for sure", which seems inevitable when Plan C is selected. Now, the framing of the decision problem suggests that the agents would do harm, which is more morally objectionable than merely allowing it. This difference explains why, in this second decision problem, agents are risk-seekers, and more individuals take the chance to avoid doing any harm, even at the risk of doing much more harm.

While this proposal amounts to a compelling answer to deflationist accounts, I still have to defend it against Kamm's position, which argues that we can get rid of framing effects upon careful scrutiny. In short, Kamm could argue that agents might be sensitive to baseline shifts and loss/no gain characterisations, and be affected by different risk attitudes; these features could, arguably, cloud their judgements about doing/allowing classifications. Nonetheless, Kamm argues, "the fact that lay experimental subjects are tricked by framing effects into identifying the baseline from which to judge losses and no-gains does not gainsay the possibility that moral theory could use the distinction in an unconfused way."[30] I thus have to answer

[30] Kamm (2007), p. 432.

Kamm's objection that doing/allowing classifications are not *ultimately* affected by framing effects.

One's response to the above claim depends on the role one assigns to intuitions in a doing/allowing distinction account. Kamm, as I explain in section 4.4, defines both decision problems in the Asian flu case as instances of "allowing harm". Arguably, this conclusion relies on a prior definition of the distinction. Specifically, what she seems to have in mind is her "imposing" intuition, which is that a conduct classifies as "doing harm" if it is intruding or imposing on another individual's body or possessions, while it is just an instance of "allowing" if this condition does not hold when bringing about the harm. In this sense, whichever vaccination plan we choose, we would not impose on these people and thus we would not be doing any harm. Note that, however, the first point in the "hidden agenda" was that people's doing/allowing classifications are largely motivated by their intuitions. Examples like the Asian flu amount to problematic cases only as long as one thinks that these seemingly inconsistent intuitions have a direct bearing on the doing/allowing distinction. Kamm's argument that our intuitions are mistaken, on the other hand, seems to adopt a different perspective on moral theorising, one that, in some sense, assigns a lesser role to our intuitions.

To be fair, Kamm could answer this point by noting that she is not claiming that moral intuitions do not play *any* role whatsoever in building an adequate account of the doing/allowing distinction. In a few problematic cases, however, at least some of our intuitions could be wrong, and not particularly reliable to serve this task. I concede this point to Kamm, even if I am worried that she is relying on a prior "correct" characterisation of doing and allowing so as to discriminate between more and less reliable moral intuitions.

Even so, Kamm's proposal downplays the import of preference reversals and inconsistent intuitions in scenarios affected by framing. In the Asian flu example, there is at least disagreement over the classification of the second decision problem as "allowing". To keep Kamm's framework, therefore, we would be forced to dismiss a lot of intuitions as wrong and misleading. Cases of framing effects, moreover, seem to be pervasive, showing a general lack of consensus and an instability in people's classifications, and it is up for debate whether they can actually be "overcome" by careful scrutiny. As persistent disagreement over the correct doing and allowing characterisations seems to be a prominent feature of many cases of framing effects, we might be interested in an account of the doing/allowing distinction that explains the experimental findings, rather than one that solves the issue by claiming that one frame is more "correct" than the other. In this respect, my hypothesis has the advantage of tracking how the reframing of the decision problem affects the perception of the same behaviour as doing rather than allowing.

To sum up, my position is consistent with Kamm's claim (a) that the doing/allowing distinction does not collapse into the loss/no gain distinction, where the latter is seen as a cognitive bias. We both agree that there is more in the "Asian flu" case than a different perception of outcomes as losses or no gains, as the focus is also on the way the outcome is brought about. I do not think, however, that Kamm succeeds in proving (b), i.e., that the doing/allowing distinction is not subject to framing. The fact that the doing/allowing distinction is different from the loss/no gain distinction does not in itself rule out that the former could still be subject to framing. The latter hypothesis, I argue, is compatible with experimental findings and acknowledges for the instability of our intuitions and the persistency of preference reversal and interpersonal disagreement.

Arguably, evidence that the doing/allowing distinction depends on how the decision problem is framed does not rule out the possibility that, with these characterisations, we might be tracking something morally relevant and significant in the way agents bring about a harm. Nonetheless, these results also amount to evidence that the doing/allowing distinction seems to be ultimately tied with and dependent on other normative and descriptive features of the decision context. In this sense, the whole project of separating the descriptive issue of distinguishing doing/allowings from the normative issue of investigating the moral significance of this distinction might prove unfeasible.

4.6 Conclusion

The pervasiveness of framing effects, which has been widely documented in the empirical literature, raises some crucial questions about the moral significance of the doing/allowing distinction. Specifically, evidence of framing effects may seriously undermine the reliability of the insight that "doing is worse than allowing" and suggest that our moral intuitions about doing/allowing are seemingly flawed. This, in turn, would be bad news for the positive thesis. So as to further support this sceptical conclusion, it has been argued that, indeed, the doing/allowing distinction can be expressed in terms that are morally irrelevant, as the result of cognitive attitudes and reasoning biases, such as the loss/no gain effect.

This conclusion can, nonetheless, still be challenged by advocates of the positive thesis, by downplaying the confounding impact framing effects have on our intuitions: Kamm, for instance, argues that the fact that our moral reasoning has some significant limitations does not prove that the doing/allowing distinction is not tracking a morally relevant feature of human behaviour. In this sense, cases like the Asian flu simply amount to particularly difficult examples where we would need to exercise our moral skills more carefully and put our first-hand intuitions under careful scrutiny. To sum up, we can concede that framing effects do not alone provide a decisive argument in favour of or against the positive thesis.

My contention, however, is that at least some examples of moral framing, such as the Asian flu case, do suggest that our doing and allowing characterisations, and our moral evaluations of conducts so described, are affected by how the decision problem is described and presented to the agent. Consistent with my discussion in Chapter 2, I conclude that doing/allowing classifications depend on framing because they depend on how one analyses the distinction in the first place. The project of pulling apart the *descriptive issue* of "where to draw the line" and the *normative issue* of whether the line tracks something morally relevant, therefore, does not look like a promising way of analysing our use of this distinction. In conclusion, both the failure of the strategy of fully-equalized cases and evidence of framing effects seem to show that we cannot disentangle doing and allowing from other descriptive and normative features of the decision context.

5. The alternative thesis

The difficulty in drawing and reaching agreement over fully-equalized cases, together with the doubts raised by the persistence of framing effects, gives us reasons to be sceptical about the positive thesis regarding the moral significance of the doing/allowing distinction. These considerations, nonetheless, do not seem to amount to enough evidence in favour of accepting the negative thesis either, that is, the hypothesis that the doing/allowing distinction is not morally relevant. In this chapter, I argue that neither of these positions can be definitively defended against all objections, and they are ultimately underdetermined by our ordinary use of the doing/allowing distinction for moral purposes. In light of this, I suggest that an alternative explanation appears very attractive. Specifically, I claim that it could be the case that there *is* something morally relevant to the doing/allowing distinction; at the same time, this distinction, despite its moral relevance, remains subject to framing, thus allowing for disagreement and instability. If we accept my proposed interpretation, it follows that the way we use the doing/allowing distinction may be messier and more ambiguous than some would like. There is something initially uncomfortable about my suggestion, as we aim to structure and make coherent our moral reality through moral theorising. At the same time, I believe that, on reflection, our use of this distinction is complex and nuanced, and disagreement is persistent. My suggestion, therefore, may track our moral reality as it is, at least in this respect.

Throughout this chapter, I defend this interpretation by comparing it to the positive and the negative theses, and at a finer level of detail, comparing it to the main categories of explanatory (or what I call *reductionist*) accounts. I conclude that my alternative thesis provides some explanatory advantages. In section 5.1, I argue that neither the positive nor the negative thesis emerges as a clear winner, and I give the first formulation of the alternative thesis. In section 5.2, I examine how this proposal preserves some valuable intuitions from both positions. To do so, I describe more accurately my proposal with respect to the import of framing effects and to the understanding of "moral relevance" (5.2.1). I then compare the alternative thesis with three main families of explanatory accounts of the doing/allowing distinction, and I outline more systematically three desiderata for a convincing framework of this dichotomy (5.2.2). In section 5.3, I elaborate on these desiderata. I show how Hitchcock's self-contained network model seems to capture all these features, that is, causal structure (i), frame-dependency (ii), and moral relevance (iii). I conclude in 5.4 with a final formulation of the alternative thesis which incorporates Hitchcock's self-contained network account as a tool for distinguishing between "doing" and "allowing" behaviours.

5.1 The positive, the negative, and the alternative theses

Returning to the formulations of the negative and positive theses regarding the moral relevance of the doing/allowing distinction, it is important at this point to be more specific about the differences between the two positions.

> **Positive thesis**: the doing/allowing distinction amounts to an unambiguous distinction, that is, there is a right (frame-independent) characterisation of actions as instances of doing or allowing respectively. This distinction is independent of framing and morally relevant.
>
> **Negative thesis**: the doing/allowing distinction might be ambiguous or unambiguous, that is, there might or might not be a correct characterisation of actions. Whichever the case, this distinction is dependent on framing (in imperfect deliberators) and it is not morally relevant, even in its right characterisation (if it exists).

In Chapters 3 and 4, I examined two main arguments which are employed in the moral literature to provide an answer to the question of whether the doing/allowing distinction is morally relevant, and, more generally, to support the positive or the negative thesis. Specifically, fully-equalized cases are used as evidence for or against the positive thesis. Evidence of framing effects, on the other hand, is mostly used by supporters of the negative thesis. These arguments, nonetheless, do not appear to be particularly striking or decisive. Regarding the attempt to come up with fully-equalized cases, and to isolate the contribution of the doing/allowing dichotomy to our moral evaluation of an action, I show that there is disagreement over allegedly fully-equalized cases. Furthermore, there are reasons to be sceptical over the feasibility of the strategy itself. Most examples, in fact, turn out to be convoluted, or to have extreme features, such that there is reason to doubt the associated moral intuitions. With regard to the persistence of framing effects, the fact that doing/allowing classifications are influenced by apparently morally irrelevant aspects raises reasonable doubts over the moral significance of the distinction.

Nonetheless, any amount of interpersonal disagreement, or evidence of "instability", can be in principle accommodated within a positive framework. Ultimately, advocates of the positive thesis could argue that there is a fact of the matter whether an action is an instance of doing rather than allowing; yet, we may often be unsuccessful in recognizing the right characterisation, due to cognitive limitations and/or particularly intricate scenarios. Moreover, attempts to explain the doing/allowing distinction in purely non-moral terms (as, for instance, in Horowitz's account) would require us to dismiss as mistaken the moral intuitions we have when comparing doings with allowings behaviours. As McMahan argues, challenging these positions:

> "If the argument is correct, how can we account for the fact that we have been systematically misled into believing that there is a moral asymmetry between making and allowing? The idea that there is such asymmetry is ubiquitous. To the best of my knowledge, there are no societies, no culture, past or present, in which failing to prevent a harm is regarded as morally equivalent, other things being equal, to killing a person (...). How did they manage to be so obtuse?"[1]

In summary, both the negative and the positive thesis appear to be *underdetermined*, insofar as all counterexamples, evidence of disagreement, and appeal to intuitions could be potentially explained and justified by both these opposing positions regarding the moral relevance of the doing/allowing distinction. Also, as I will explain in 5.2, both the positive and the negative theses seem to require us to sacrifice some of our immediate intuitions, and to recognize that our moral reasoning skills are significantly flawed in many circumstances. Supporters of the positive thesis take our case-based intuitions about the asymmetry of the distinction more seriously. Supporters of the negative thesis, on the other hand, prize more coherence in theorising.

The state of the literature thus gives us a reasonable justification for exploring different interpretations and alternative ways of thinking about the doing/allowing debate. Specifically, rather than continuing to focus on how we can isolate and examine the moral relevance of the distinction, we may ask why this distinction seems to be particularity resistant to analysis, and ultimately embedded with other moral principles and classifications.

In this chapter, I start outlining such an alternative thesis, which I provisionally define as follows:

> **Alternative thesis**: the doing/allowing distinction is inherently ambiguous, that is, there is not always a fact of the matter whether an action is an instance of doing rather than allowing; classifications of the same action as the former or the latter can be descriptively adequate, depending on the specific framing. The fact that an action is described as doing rather than allowing is nonetheless morally relevant.

Note that by moral relevance I do not mean here that all instances of doing harm are morally worse than all instances of allowing harm, nor that doing harm is always impermissible while allowing harm is always permissible. As Woollard (2015) suggests,[2] the asymmetry of the doing/allowing distinction is better interpreted as a claim about justification: *all other things*

[1] McMahan (1998), p. 397.

[2] Woollard (2015), pp. 8–9.

being equal, doing harm is harder to justify than allowing harm to occur.

In defending this hypothesis, I will move away from constructing cases that support the negative, the positive, or the alternative thesis. I think that a way of arguing in favour of my proposal, which advances the discussion, is to focus on its explanatory merits. Specifically, in my view, disagreement between the positive and the negative thesis is so pervasive because they both capture aspects of the doing/allowing distinction we strongly feel are right. In the following section, I will argue that the alternative thesis in fact allows us to keep these compelling aspects of both the positive and the negative theses, and that this amounts to an attractive feature of this proposal.

5.2 Defending the alternative thesis

We have seen that both the positive and negative thesis imply that there is something faulty about our moral intuitions: in the case of the former, it is the biases of framing, and in the case of the later, it is the mistake in taking a distinction to be morally relevant when it is not. This raises the question of how our moral reasoning should respond to intuitions.

My contention is that the alternative thesis helps us out of this impasse. Specifically, the alternative thesis preserves two important aspects of both theses: the intuition that the doing/allowing distinction is morally significant and the fact that it might be context-dependent. Note that, while this project may seem doomed, an attempt to "square the circle", it in fact exemplifies nuanced moral theorising in response to intuitions. We start with our intuitions over specific examples, try to make sense of them and treat them as uncovering the underlying principles explaining and justifying these intuitions. During this process of analysis, however, we might find out that different counterexamples and intuitions apparently contradict our first-hand insights, and that our contradictory intuitions resist a coherent systematization. We must thus select which intuitions to keep and which to abandon as inconsistent and unsupported, and we strive to find a balance in such operation. As McMahan (2000) puts it,

> "One of the aims of moral theory is to illuminate the considerations that underlie our common moral intuitions. Yet it may happen that these deeper considerations, when exposed, seem not to be especially cogent or compelling. When this happens, we face a choice between retaining intuitions that are apparently ungrounded and abandoning them. Yet the intuitions may be central to any morality that we could bring ourselves to accept-indeed any system of norms that we could genuinely recognize as morality at all. I think it is

> possible that a dilemma of this sort arises with our intuitions about killing and letting die."[3]

What McMahan describes here is a method now known in the literature as *reflective equilibrium*, a term coined by Rawls in "A Theory of Justice". While I do not aim specifically to replicate this argumentative strategy, my goal in presenting the alternative thesis is to find a "balance" with respect to both the positive and the negative theses.

In 5.2.1, I further elaborate on how the alternative thesis walks a desirable line between the negative and the positive thesis. In 5.2.2, I compare the alternative thesis with other explanatory accounts of the doing/allowing distinction, and outline three main desiderata for a suitable framework of this dichotomy.

5.2.1 The alternative thesis in the "space" of positions

In the discussion above, I argued that the alternative thesis may successfully account for the fact that the doing/allowing distinction is morally significant, *and* for the fact that two classifications of the same action as doing or else allowing can be legitimate, thus explaining disagreement and framing. *Prima facie*, these two aspects might look inconsistent. To see how the alternative thesis can, in fact, accommodate both, it is useful to offer a more fine-grained taxonomy of the positions one could take with respect to the moral relevance of the doing/allowing distinction.

Note that, when we talk about the moral relevance of the doing/allowing distinction, there are at least two different issues under investigation, which can be summarized as two separate questions. In this sense, the "space" of the discussion can be arranged along two axes, which appeal to these two different questions about moral relevance. Opposite answers to these two questions determine different positions. The first question addresses the stance towards moral framing:

1) Do framing effects obstruct moral truth?

In other words, this question asks whether what is morally relevant can include aspects which are associated with frames, or whether frame-dependency rules out moral significance.

Note that, with frame-dependency, I mean here "genuine" frame-dependency rather than frame-dependency that is obviously due to mere cognitive mistakes.

The second question, on the other hand, appeals to the conditions for determining whether the doing/allowing distinction is morally relevant:

[3] McMahan (2000), p. 110.

2) Is the doing/allowing distinction morally relevant *per se*?

This second question, which I will examine in more detail in 5.2.2, considers the issue whether the doing/allowing distinction can be successfully reduced to other features or principles, or rather amounts to an independent characteristic of actions.

Now, defenders of the positive thesis would answer "yes" to both questions: framing effects obstruct moral truth, are misleading, and need to be corrected when examining the doing/allowing distinction (as they are simply cases of bad human reasoning); moreover, the moral significance of the distinction cannot be further reduced to independent principles or characterisations. What about the negative thesis? To be sure, the answer to the first question will still be "yes": some supporters of the negative thesis would argue that the doing/allowing distinction is, in fact, ultimately subject to framing, and thus cannot be morally significant. Regarding the second question, remember that the negative thesis, as first formulated in Chapter 3, p. 55, only argues that the doing/allowing distinction is not morally relevant *per se*. In general, all proposals which could be classified in the negative thesis field claim that the doing/allowing distinction ultimately tracks other features or principles. Deflationist interpretations further argue that these other features are morally irrelevant, while what we might dub the "moderate negative thesis" seems to allow for the possibility that these separate principles are morally significant. So, for the negative thesis, the answer to the second question is "no", but on some accounts the doing/allowing distinction nonetheless maps onto morally relevant classifications, and thus retains moral significance, in spite of framing effects.

I can now also define my proposal more accurately. First, the alternative thesis argues that frame-dependency does not rule out moral significance, so the answer to the first question is "no", unlike both positive and negative theses. That is, there could be no fact of the matter whether an action is an instance of doing rather than allowing, as these classifications can be context- and agent-dependent. Second, the alternative thesis merely states that doing/allowing characterisations should be relevant to the moral evaluation of cases, without further arguing whether it is morally relevant *per se* or rather tracks other morally relevant principles and features. In 5.3 and in Chapter 6, I will elaborate on this point, arguing that doing/allowing classifications might not be morally significant *per se*, but the way we make these classifications incorporates morally relevant aspects. These moral features, I will argue, cannot, however, be easily systematized with a unitary explanation, that is, one which identifies a fixed set of principles or characteristics that the doing/allowing distinction univocally maps onto. In this sense, the answer to this second question might be ambiguous. On the one hand, the doing/allowing distinction is morally relevant because it incorporates other morally relevant features, so, in this respect, it is not morally significant *per se*. On the other, the distinction is not identical to (it does not univocally map onto) any other independent moral categorisation. I note that this might be a merely terminological issue, and one I will not explore here. I will thus leave it open to one's account of what counts as “morally relevant*per se*”.

To sum up, the above discussion can be summarized in Table 5.1 as follows:

		The doing/allowing distinction is morally relevant *per se*	
		Yes	No
Framing effects undermine moral relevance	Yes	Positive thesis	Negative thesis
	No	Alternative thesis	

Table 5.1

Bearing these classifications in mind, we can further proceed in refining the alternative thesis, and in examining its explanatory merits. To do so, I discuss this position with respect to three main families of explanatory accounts. I argue that these approaches draw attention to some crucial aspects of the doing/allowing distinction, which I use to outline three desiderata for a convincing account of the doing/allowing distinction. Specifically, apart from moral relevance and frame-dependency, I argue that we should also account for the intuition that the distinction tracks an agent's causal contribution to an outcome.

5.2.2 The alternative thesis, explanatory proposals, and three desiderata

This section sets out to examine in a more systematic way the explanatory advantages of the alternative thesis. To do so, I first define the alternative thesis as a *reductionist* account. This conclusion, I suggest, is justified by the difficulties surveyed in Chapters 3 and 4 in disentangling the doing/allowing distinction from any other feature of the context. I then compare my proposal with other reductionist/explanatory accounts of the doing/allowing distinction. My aim is to argue that, while these different families of frameworks seem each to fit particularly well with one specific aspect of the doing/allowing distinction, none of them accounts for all these features together. Specifically I identify three main desiderata, which the alternative thesis satisfies at the same time.

In the section above, I have been talking about accounts which argue that the doing/allowing distinction can be expressed by other principles and features in a fairly informal way. One way to define such accounts might be the label “reductionist”. Simply put, I distinguish between *reductionist* and *non-reductionist* accounts of the doing/allowing distinction by appealing to how they answer to the following question: **is the doing/allowing distinction a fundamental one, or can it be expressed by other principles or features?**

I define as *reductionist* all accounts which claim that the doing/allowing distinction is *not* a fundamental one. Which position should the alternative thesis take? As I concluded in 4.6, the failure of the strategy of fully-equalized cases, and the import of framing effects, seem to show that the doing/allowing distinction cannot be easily disentangled from normative and descriptive features of cases. For this reason, I think the doing/allowing distinction might be most promisingly cashed out in a way which incorporates other principles and features.

I now discuss which “features or principles” seem to be on the table, and could suitably serve a reductionist. To do so, I attempt a taxonomy of reductionist accounts of the doing/allowing distinction. This task will help in singling out some key aspects of the distinction, each captured by one of these reductionist proposals. Two of these features (moral relevance and frame-dependency/ disagreement) have already emerged in the comparison between the positive and the negative thesis; one additional feature will define a list of three desiderata. As it will become clear from my discussion, I do think that most of the accounts of the doing/allowing distinction are, in fact, reductionist. In this sense, this survey is crucial for understanding in which way the alternative thesis is reductionist.

A first class of reductionist models aims to express the doing/allowing distinction in terms of features which are not, at least *prima facie*, morally relevant. These kinds of accounts can be further divided into *cognitive* and *descriptive* interpretations. The former argue that the doing/allowing distinction can be fully explained as the result of cognitive biases, or, more charitably, of the characteristic way our moral reasoning works. These models, like Horowitz's or Sinnott-Armstrong's accounts, have the merit of allowing for framing effects, and, at the same time, of explaining why the doing/allowing distinction amounts to a strong and persistent intuition, as it is the product of the exercise of our (flawed) reasoning skills. Descriptive models like Bennett's account, on the other hand, explain the doing/allowing distinction as capturing some factual characteristic of actions, such as being positively or negatively relevant to an upshot or, more generally, as being relevant to an outcome in a specific way. After identifying such features, they usually deny that they are morally relevant.

This class of causal frameworks seems to adequately identify one aspect which is crucial for our doing/allowing attributions, that is, the fact that an agent has a specific type of causal relation with the outcome. Purely causal accounts, nonetheless, appear at least incomplete, as causes are always defined against a set of background conditions, such as what the natural course of nature is, or what the best or standard explanation for an outcome is. As I argued in Chapter 2, these models thus rely on prior evaluations and expectations about the context, and should be supplemented with a precise interpretation of how this background process works. In this sense, these accounts are somehow naïve in expecting that causal facts are "out there", independently of our interpretation of the situation.

Both the cognitive and descriptive models, in summary, reduce the doing/allowing distinction to seemingly non-moral features of actions: the idiosyncrasies and bugs of human cognition, or some merely descriptive facts about an action, which do not apparently amount to morally relevant characteristics. The result of these efforts, as Bennett and other reductionists argue, thus forces us to question and possibly abandon a particularly strong and long-standing intuition, and to accept that we are mistaken in attaching moral significance to doing/allowing characterisations. To be clear, causal accounts do not necessarily need to deny that the doing/allowing distinction is morally relevant. Some of the "sequence" accounts I surveyed in Chapter 1expemplify this possibility. What I mean here is that causal facts are not *obviously* morally relevant.

Reductionist proposals, nonetheless, do not necessarily need to look like Horowitz's or Bennett's account. There are proposals that may be classed as versions of what I have called "the moderate negative thesis", that is, accounts which argue that the doing/allowing distinction is not morally relevant *per se*, but cannot be fully expressed in terms that are morally irrelevant. Specifically, one can argue that the doing/allowing distinction is not a fundamental one, but rather it tracks a separate set of moral principles or categorisations. One example of this strategy is Quinn's account, which defines doing and allowing as negative and positive rights violations. Another reductionist approach that belongs to this class is Kagan's norm-violation proposal, which identifies doing actions as violating a rule of behaviour. As I argued in Chapter 2, these accounts often fall short in accounting for all our use of doing/allowing classifications, especially in less "moralised" examples.

Note that, at this point, one might argue that all explanatory accounts of the doing/allowing distinction could be defined as reductionist, insofar as they provide an explanation of this dichotomy referring to *something else*. In this sense, most frameworks in the "positive thesis" field would also fall into the third reductionism category, insofar as they refer to other moral principles, like Foot's self-ownership framework. Clearly, this would mean that the term "reductionist" is redundant, as I would label as such any explanatory attempt. Only frameworks which argue that the difference in moral worth of doing and allowing actions is somehow "self-evident", and does not require further justifications, would indeed truly characterise this dichotomy as a fundamental one.[4] While this aspect of my classification might appear problematic, I see it as the result of the difficulties illustrated in the opening of this chapter, which any doing/allowing account must confront. The problems raised by our intuitive use of the doing/allowing distinction do, arguably, require a kind of analysis which cannot be easily dismissed within a "self-evident" framework. For taxonomy purposes, however, it is more natural to define as "explanatory accounts" all frameworks which provide a complex and not self-evident model of the distinction. Within this broad category, we can keep the label "reductionist accounts" to refer to deflationist frameworks, like Horowitz's or Bennett's, which deny the moral relevance of the distinction. Note also that, in this respect, what is more informative is not whether an account is explanatory/reductionist, but rather *in which way* it is reductionist. In the remaining of this thesis, and in 5.5 more specifically, I will try to answer this question regarding the alternative thesis.

In summary, we can identify three main explanatory accounts, which amount to human cognition models (reductionists), descriptive/causal models (reductionists), and models appealing to other moral features (non reductionist in the narrow sense). All these proposals, I have argued, seem to fit particularly well with a key feature of the doing/allowing distinction, but also fall short in accounting for other fundamental aspects. Specifically, I think that each of these three families of accounts identifies a crucial characteristic of this classification: (i) the fact that doing/allowing classifications track whether an agent is causally relevant to the outcome in a certain way, (ii) the fact that doing/allowing classifications can be frame-dependent, and (iii) the intuition that the doing/allowing distinction is obviously morally relevant. Table 5.2 summarizes the discussion above, showing which explanatory account

[4] One may argue that, after all, moral principles do not always require an external justification, in terms of independent considerations or rules: it is sound, for example, to think that there is no need to motivate a moral proposition like "stealing is bad" or "being loyal is good". Frameworks appealing to ethical intuitionism, for example, typically argue that such basic moral propositions are intrinsically incapable of proof. The doing/allowing distinction, in this sense, could simply amount to one of these basic principles, and the insight that "doing is worse than allowing" would not call for more discussion, as all moral agents would simply be able to recognize and employ it in appropriate contexts. I also note that it is not only intuitionists in the Moorean sense who can see the dichotomy as fundamental. Jeff McMahan's and Frances Kamm's accounts might be "fundamentalists" too, in the sense that they would define the distinction, and then, through the use of many cases, show that it matters to our intuitive judgments, and that it coheres well with other judgments we make. The relevance of the distinction for such accounts is not self-evident; it is justified through a reflective equilibrium process, but there is no "explanation" as to why it matters that could be characterised as reductionist.

matches each of the three aspects of the doing/allowing distinction:

	Tracking that doing and allowing are "causing" in a specific way	Allowing framing effects	Obvious moral relevance
Human cognition models		✓	
Descriptive models	✓		
"Other moral principles" models			✓

Table 5.2

I think that i), ii) and iii) amount to three reasonable *desiderata* for a suitable explanatory account of the doing/allowing distinction. My contention is that the alternative thesis, with respect to the three models above, accounts for all these features. In the remainder of this chapter, I justify this claim.

In section 5.3, I argue specifically that the alternative thesis may be most promisingly formulated using the self-contained network account. This formulation can conveniently captures (i), (ii) and (iii).

5.3 The self-contained network account and the alternative thesis

In Chapter 2, I argued that "mixed" accounts seem particularly well-equipped to track our use of the doing/allowing distinction, and I examined how we can employ Hitchcock's self-contained network framework to justify most of the ordinary doing/allowing attributions. In this section, I show how this model also allows for doing/allowing attributions to be frame-dependent, thus incorporating framing effects in this account. Arguably, this solution accounts for all three desiderata, as we would have a reliable descriptive model which is causal in nature (i), explains framing effects (ii), and grounds doing/allowing attributions in morally relevant aspects, thus capturing the moral significance we attach to this distinction (iii). I therefore suggest that a formulation of the alternative thesis which incorporates this model will nicely retain these explanatory merits.

As I argued in Chapter 1, the key “gap” of causal accounts of the doing/allowing distinction is the assumption of which course of events, or “sequence”, is considered to be natural or relevant. The issue of the dependence on the “normal course of events” is involved in distinguishing “true” causes from background conditions and in distinguishing acts from causally relevant omissions. While discussing Hitchcock's self-contained networks proposal, I suggested that normative considerations could help us fill this gap. Expectations over which behaviours are standard, which duties are morally required, and which norms should not be violated define which kind of causal relations are relevant and salient to us. Just like the presence of oxygen is a standard background condition, and thus it is the lighting of the match which causes the fire, so feeding one's baby is a moral requirement for all parents, and therefore failure to fulfil this duty is the salient cause of the death of the baby. Within Hitchcock's model, this substantive work is done at the level of setting the default and deviant value of variables. To be sure, the core of this model still captures “causal facts”, that is, the “child” variables in the model counterfactually depend on the “parent” variables. Whether an action counts as doing rather than allowing, however, depends on whether the causal network is self-contained, a parameter which is settled by the assignment of deviant and default values to variables.

In 5.3.1 and 5.3.2, I show in detail how this model allows for shifts in frames to change the characterisation of actions as doings rather than allowings. Also, this account makes clear that attributions of doing and allowing capture and express morally significant considerations, as doing/allowing attributions are built upon a normative background. I will examine this second aspect in 5.3.3 and in greater detail in Chapter 6.

5.3.1 Incorporating framing effects

Consider again one of the most paradigmatic cases of framing effects, the “Asian flu” experiment analysed in Chapter 4. I have previously suggested, when discussing Kahneman and Tversky's interpretation, that the use of specific phrasing like “200 people will be saved” rather than “400 people will die” could induce a different perception of the same vaccination plan as an instance of doing harm rather than allowing harm to occur. But how does this work in more detail? In Chapter 4, my observations were limited to the fact that setting the baseline at “everyone lives” rather than “everyone dies” does the trick. I now suggest that we can convincingly cash out this idea at the level of default value assignments to variables.

Recall that, in the “Asian flu” case, agents are faced with the threat of an “Asian flu” which is expected to kill 600 people, and can choose between two alternative vaccination programs:

- If Program A is adopted, 200 people out of the 600 will be saved.
- If Program B is adopted, there is 2/3 probability that no-one will be saved, and 1/3 probability that all 600 people will be saved.

Alternatively, in the second experimental setting, agents are asked to choose between C and D:

- If Program C is adopted, 400 people out of the 600 will die.
- If Program D is adopted, there is 1/3 probability that no-one will die, and 2/3 probability that 600 people will die.

In Chapter 4, I argued that preference reversal in this case can be explained by the fact that, while choosing Program A is characterised as allowing harm, choosing Plan C is characterised as doing harm, despite these two programs being equivalent in terms of expected lives saved (or lost). In this sense, the shift in the classification of the same action as an instance of doing or allowing is a straightforward example of a framing effect, as it is merely motivated by the choice of specific words, a seemingly irrelevant modification.

Hitchcock's causal model allows, in this case, for a consistent reconstruction of the shift in classification. Recall now the main definitions of the Hitchcock framework: a *causal model* is an ordered pair <V, E>, where V is a set of variables and E is a set of equations among these variables. In my simplified interpretation, variable can take two different values, where each value represents the occurrence (or non-occurrence) of some event. By considering Plan A and Plan C in isolation, without referring to the comparison with the risky options (Plan C and Plan D respectively), we can thus define the following variables:

A = 1 if the agent selects plan A, A = 0 if she doesn't;

C = 1 if the agent selects plan C, C = 0 if she doesn't.

In terms of the outcomes, the more “natural” interpretation would be to set O (the outcome) as 200 lives saved and 400 lives lost. Nonetheless, as we are focusing here on the “harm” the programs bring about, I narrow my analysis to the 400 lives lost, and on which type of causal connection the agent has with these 400 lives lost. Therefore, O = 400 lives lost.

The equations representing the counterfactuals we use for describing the model, are:

A = 1, then O = A (if Program A is selected, 400 lives will be lost);

C = 1, then O = C (if Program C is selected, 400 lives will be lost).

Hitchcock's notion of counterfactual dependence can be represented, with respect to the Asian flu case, by the following graphs:

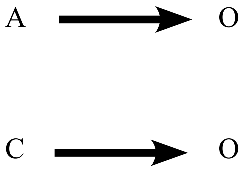


Figure 5.1: Asian flu

Where A is a parent of O in the first causal model and C is a parent of O in the second model.[5]

In both cases, O counterfactually depends on A or C (respectively), as the counterfactuals “if Program A had been selected, there would have been 400 lives lost” and “if Program C had been selected, there would have been 400 lives lost” are both true. The “negative” counterfactuals are also true: “if A had not been selected, there would not have been 400 lives lost”; “if C had not been selected, there would not have been 400 lives lost”.[6]

Let's now consider whether A and C count as doings rather than allowings. As per my discussion of Hitchcock's model, this depends on whether the causal networks {A, O} and {C, O} are self-contained. According to Hitchcock, if <V, E> is a causal model, X, Y ∈ V, and N ⊆ V is the causal network connecting X to Y in <V, E>, then the causal network N is self-contained if and only if for all Z in N, if Z has parents in N, then Z takes a default value when all of its parents in N do (and its parents in V\N take their actual values). Conversely, a causal network is non-self-contained if it is possible for Z to take its deviant value while all its parents take their default ones. Following this distinction, I further defined cases of doing as instances where the outcome counterfactually depends on the action, and the network is self-contained. On the other hand, I defined cases of allowing as instances where the outcome counterfactually depends on the action, and the causal network is non-self-contained.[7]

What we need to find out, therefore, is whether O takes a default or deviant value when its parents A and C take their default value. To settle this point, I suggest that we should consider that agents' expectations about which is the default value of O might be affected by the change in the baseline. For this reason, I define O* as the outcome in the first decision problem, and O' as the outcome in the second decision problem.

For the first decision problem, recall that the framing in terms of “saving” seems to imply that

[5] I am ignoring here the causal contribution of the flu.

[6] To see why the negative counterfactuals are also true, we may think that, if Programs A or C are not selected (and, consequently, B and D are selected) these 400 people would instead have a 2/3 chance of dying, which is arguably a different outcome.

[7] I also argue, following Hitchcock, that “real” instances of allowing will be the ones where the variable is not a mere background condition, and thus takes its deviant value. I do not dwell on this further condition here, as I focus narrowly on distinguishing doing from allowings, rather than “real” allowings from background conditions.

without the intervention of the agent everyone will die. Therefore:

O* = 1 if 400 lives are not lost, O* = 0 if 400 lives are lost,

where Def(O*) = 0.

Note that I set here the default value of O* as 0 because, if the natural course of events seems to be that everyone will die, the expectation is that these 400 people will die anyway. Following Hitchcock's suggestion,[8] I also set Def(A) and Def(C) = 0. In this case, it is possible for, O* to take its deviant value of 1 when A takes its default one of 1, that is, 400 lives are saved when the agent does not select Program A. This happens exactly when Program B is selected, and nobody dies. The causal network {A, O*} can be thus defined as non-self-contained. This represents how Program A is characterised as allowing harm to 400 people.

As I argue in Chapter 4, the use of the phrase "saving" is what induces agents to think at the baseline, and thus set the default of O* as "everyone dies". But what about Program C? In the second decision problem, the baseline is set at "everyone lives". We can thus define:

O' = 1 if 400 lives are not lost, O' = 0 if 400 lives are lost,

where Def(O') = 1.

Note that I set here the Def(O') as 1: if the "normal course of events" is perceived as everyone lives, the expectation is that the 400 lives will not be lost. Again, the default value for C is set at 0. In this case, it is not possible for O' to take its deviant value (that is, that not everybody lives) when C takes its default one, that is, if Program C is selected.[9] The causal network {C, O'} is therefore self-contained, and this would explain the characterisation of the choice of Program C as doing harm.

To refer to Hitchcock's understanding of self-contained networks, I suggest that the first decision problem stresses the role of the flu, and thus makes us perceive the selection of Program A as an incomplete explanation of the outcome. On the other hand, in the second decision problem, the import of the flu is concealed, and thus the selection of Program C is perceived as a complete explanation of the outcome.

The setting of default and deviant variables can thus potentially explain why the same action – in terms of expected lives lost – can be classified as doing rather than allowing depending on how the description of the problem influences the choice of default/deviant. Similarly, the

[8] Hitchcock (2007, p. 507) argues that "In the case of human actions, we tend to think of those states requiring voluntary bodily motion as deviants and those compatible with lack of motion as defaults."

[9] More precisely, it is only possible for O' to take its default value of 1 if Program C is not selected, and thus takes its default value as well: this would be the case where Program D is selected instead, and everyone lives.

model can also explain why agents may disagree about such classifications. Some readers, for instance, might not be persuaded by the reconstruction above, and argue that, in the causal model {C, O'}, the default value of O' should be still put at 0. For these agents, opting for Program C would still be characterised as allowing harm.

Of course, this conclusion, which can be generalised for most cases of frame-dependency and disagreement, is not particularly comforting in the light of our aspiration to reach agreement and be consistent. Nonetheless, as I will argue later in this section, not *all* doing/allowing characterisations are subject to such ambiguity and, more significantly, in less controversial and complex cases, agents will be less justified in setting alternative default values. That is to say, sometimes there will be a "correct", or at least robust, characterisation of actions as doings or else allowing. In examples characterised by persistent interpersonal disagreement, however, two different value assignments to a variable seem plausible. In such circumstances, shifts in frames will give agents additional reasons for setting the default value in one way rather than the other.

5.3.2 Framing: empirical and normative expectations

In my formulation of the alternative thesis, I rejected the position that there is a "correct" characterisation of actions. In the Asian flu case, we can thus argue that, both the "A" characterisation as allowing and the "C" characterisation as doing are tenable. One thing, however, is to represent how different doing/allowing classifications can be brought about by different framings; another is to argue that different framings give rise to different moral contexts in which we intuitively apply different moral rules. In this respect, the self-contained network model only shows that two alternative classifications are formally possible, but cannot provide an account of what is going on in cases of frame-dependency, and, more importantly, how this impacts the moral significance of the doing/allowing distinction.

I think that there are two routes through which framing effects can induce two different doing/allowing classifications, which map onto two ways of setting the default/deviant value, and which amount to different ways of specifying which is the normal course of events.[10] First, the "normal" course of events can be interpreted in an empirical or statistical sense, and tracks what an agent thinks is natural or more likely to happen. Different agents might have different expectations about which course of events is more standard or normal. Moreover, in addition to *inter*personal disagreement, some examples of framing effects point to *intra*personal

[10] These two different interpretations of "normal" have also been suggested by Kagan (1998), when discussing his "norm-violation" account.

disagreement, thus implying that different different frames may suggest different empirical expectations in the same agent. As I argued in the previous chapter, I do not think we have enough empirical evidence to provide a detailed explanation of this phenomenon whereby the wording of a case leads to different empirical inferences. My intuition is that, especially in the case of intrapersonal disagreement, this issue should be investigated in cognitive sciences or behavioural psychology settings. What I am more interested in here is the bearing of this mechanism on the significance of the doing/allowing distinction. When doing/allowing attributions depend on empirical expectations about which course of events is natural, or which outcome is more likely to occur, we might conclude that they capture the specific way the agent brought about the outcome. My account would thus be a purely causal one. These different ways of causing a harm might still amount to a morally relevant distinction, yet are ultimately ambiguous. As the self-contained network model captures, different empirical expectations can isolate an action as a cause or else regard it as a background condition.[11]

Second, the "natural" course of events may also refer to social rules and societal expectations. In this sense, which course of events is standard tracks agents' different expectations about what is required, or which moral rules should be relevant or more important for the case under scrutiny. In this case, disagreement can track the fact that different agents perceive different norms as salient, which may again be dependent on cognitive features. Alternatively, different default values can reflect actual moral disagreement. In this case, the doing/allowing distinction would not be morally relevant *per se*, as it captures these independent normative features, and is ultimately ambiguous. I will examine this second route in more detail in the next section, and carry on this task in Chapter 6.

Note that, while these two different routes for setting the default variable are, in principle, distinguishable, in most cases they overlap. Specifically, while we can talk about the doing/allowing distinction as an empirical matter, as in my analysis in Chapters 1 and 2 we usually think about these classifications for moral purposes. In practical instances of moral

[11] In this thesis, I do not take any specific stance with respect to the position that when the doing/allowing distinction only captures the way an agent may be causally relevant to an outcome, it still amounts to a morally relevant feature. This agnostic position, however, does not have much impact on my subsequent discussion of moral relevance. I think, indeed, that in the vast majority of cases in which we use the doing/allowing harm distinction, we do so in a "moralised" way, that is, in ethically sensitive contexts or when we want to convey moral considerations. For this reason, in practice, doing/allowing classifications usually incorporate normative features. Nonetheless, in the remainder of this study I will often use the expression "obvious" moral relevance to refer to these normative considerations, as opposed to merely "causal" considerations.
Note too that this purely empirical doing/allowing classification might capture the intuitive act/omission distinction. I do not elaborate here on this suggestion; if this were the case, however, we could conclude that the doing/allowing distinction incorporates, but expands, the act/omission distinction, in a way which is particularly evident in moral cases.

reasoning, our expectations about what is likely to happen are at least partially affected by what we think should happen, and what rules we hold each other to. Therefore, I am doubtful that there will be many cases of purely factual disagreement, in which agents' different expectations about the normal course of events will be solely dependent on different empirical expectations about which course of events is more likely to happen. The Asian flu case, in this respect, seems to be an example in which disagreement is mostly factual, and seemingly affected by some cognitive features which involve perceiving different reference points as salient.[12] Even in this case, the setting of default values may also be affected by what agents think is required from them in such ethically sensitive decisions: for instance, that they should try to avoid a situation in which someone dies for sure, or that they should not "gamble" with people's lives.

As the discussion above shows, we can account for instability and disagreement in doing/allowing classifications by appeal to the fact that there are two plausible and reasonable attributions of default values to a variable which serves as parent, and thus defines which course of events is considered "normal", both in an empirical and normative sense. This explanation gives us further elements to examine disagreement over doing and allowing attributions. It seems that many experimental results reporting instability and disagreement, and most standard cases of framing effects, refer to cases where the situation is *under-described*. In these circumstances, the fact that some details of the situations are missing could induce different agents to appeal to different normative expectations or descriptions of events, filling the picture with their intuitions about what is familiar, common and standard. Many artificial case-studies, such as the trolley case or the Asian flu case, are under-described in some respect, and people's intuition is that they "would like to know more" about these scenarios so as to make a competent evaluation.

When eliciting responses over the Asian flu case, and asking for motivations in favour of one program rather than the other, subjects often enquire about aspects of the decision problem which are not made explicit, such as "how will the 200 individuals die?", "can some of the people who get the flu survive?", "how old are these people?", "am I the person in charge?" and so on.[13] I think all the information that people want to know about by way of "setting the scene" contributes to defining which course of events is perceived as more natural or standard, and the more vague or ambiguous the description of scenarios is, the greater the case for multiple

[12] Specifically, as I will show in the following paragraphs, this example is significantly under-described: therefore, agents have no choice but to infer from clues in the wording what is the normal course of events.

[13] I am grateful to my audience in King's College, Porto, Sheffield and Utrecht, as well as to Choice Group fellows, for raising and discussing these questions. I am also grateful to my PH222 students and PhD colleagues. My anecdotal evidence on the Asian flu case builds upon these insightful contributions.

default variables being justified.

This hypothesis, moreover, seems to be supported by some experimental surveys, like the ones conducted by Jou, Shanteau and Harris (1996). The authors argue that "people have prototypical knowledge about certain types of events and comprehend the relationship between events by referring to such general knowledge structures known as *schemata.* When encountered events cannot be fit into a schema, the relationship between the events will not be understood".[14] To test for this interpretation, which is consistent with my discussion of framing effects in doing/allowing classifications, Jou, Shanteau and Harris designed some case-studies which are identical to standard examples in the experimental literature, like the Asian flu case, but contain what the authors define as a *rationale*, or causal schema. According to their hypothesis, providing this schema should help subjects to recognize the equivalence relationship between, say, 200 lives saved and 400 lives lost. So, for the Asian flu case, the following background story was provided before the evaluation of the standard options:

> "Imagine that the US has been attacked by an unusual and deadly disease. Without treatment, a person who has contracted the disease will die in a few days. Six hundred people have been diagnosed as having contracted the disease. Some substance, extracted from living human organs and extremely difficult to obtain, can cure the disease. Unfortunately, there is only enough of this substance for 200 people. No additional source of this substance will become available for at least 18 months, and no other cure or treatment will be found in at least the next two decades. If the patient receives an insufficient dose, there is a chance that the patient may live or may die. Two alternatives are proposed."[15]

The results showed that preference reversal was significantly reduced when this story was provided.[16] I do not focus here specifically on Jou, Shanteau and Harris's causal schemata proposal; what is interesting about their experiment is that the story describes in more detail the Asian flu case, and answers some of the questions discussed above which people often ask when confronted with the case. These results, in fact, seem to back up my hypothesis that different doing/allowing classifications are more likely to occur when cases are unfamiliar and significantly under-described, especially with respect to the underlying causal process.

Following the above discussion, we can thus expect that, as framing effects control the values assigned to variables, our doing/allowing classifications would be less dependent on framing in cases where the attribution of a specific default value is particularly robust. Arguably, this will happen when the scenario is very familiar to the agent making the evaluation, and/or when this

[14] Jou, Shanteau and Harris (1996), p. 2.

[15] Ibidem, p. 13.

[16] Ibidem, pp. 4–5. Specifically, the authors report that the results are statistically significant; the percentage of risk-averse responses decrease when the rationale is provided.

is made explicit in the description of the case. For instance, for those who are convinced by Kamm's reconstruction of the Asian flu case, it can be argued that the "allowing harm" description is more reasonable and natural. While it will always be possible, in principle, to explain different doing/allowing attributions using the default/deviant variables model, the justifications for setting an alternative default value will look more and more tenuous when cases are very familiar and detailed. Of course, framing effects might not be the only source of disagreement and instability in doing/allowing classifications; substantial moral disagreement may occur as well. The deviant/default variables model, I argue, also explains this phenomenon, through what I call here the "second route".

5.3.3 Normative considerations

So far, I have mostly been referring to how different empirical expectations can set different default values, like perceiving one reference point to be more natural than another in evaluating outcomes. Expectations, however, also refer to normative rather than merely descriptive features of scenarios, such as agents' intuitions about which rules of behaviour are salient, or which is the appropriate moral principle for the decision problem at issue. In these circumstances, the description of an action as an instance of doing rather than allowing incorporates normative considerations and moral intuitions, and the persistence of interpersonal disagreement can here reflect substantive moral disagreement. A straightforward example in this sense, is the controversial issue of the moral legitimacy of euthanasia. Let's take the case of active euthanasia, in which the doctor voluntarily administers a lethal drug to a terminally ill patient. I will not engage here in the broader discussion of whether this action is morally permissible or impermissible, but I focus on the limited question of whether administering the lethal drug amounts to an instance of doing harm rather than allowing harm to occur.

Within the self-contained network model, we can define the variables as follows:

A = 1 if the doctor administers the drug, A = 0 if she doesn't;

B = 1 if the patient dies, B = 0 if she doesn't.

Where B counterfactually depends on A, and A is thus the parent of B (if the doctor doesn't administer the drug, the patient doesn't die, or at least does not die at t_1).

Arguably, agents who would tend to classify this action as a case of doing harm, will characterise as "standard" a specific conduct for doctors, namely, that they should cure diseases and only intervene so as to improve patients' health conditions. In this case, recognizing this duty as the salient moral aspect of this scenario would set Def(A) = 0. On the other hand, the most natural interpretation of Def(B) is 0, at least for the death of the patient at the specific time t_1 immediately following the lethal injection. Therefore, when A takes its deviant value, and thus the doctor administers the drug, B takes its deviant value as well; if the doctor's action follows the default, and she does not administer the lethal drug, the patient does not die at time t_1; thus B takes the default value as well. We are then justified in characterising the doctor's action as an instance of doing harm, as the causal network {A, B} is self-contained.

On the other hand, agents who would tend to define active euthanasia as a case of allowing harm, would probably perceive other principles or rules of conduct, such as that the primary duty of a doctor is to act accordingly to the patient's will as more morally stringent. Therefore, Def(A) will be set at 1 when the patient asks for active euthanasia; this means that when A takes its default value, B takes its deviant one. A is then classified as allowing harm, as the causal network {A, B} is not self-contained.

In this case, the use of the self-contained network approach makes explicit that doing/allowing characterisations are *obviously* morally significant: they incorporate the fact that an agent perceives one duty/rule as more morally relevant or stringent than another. This kind of analysis thus equips us with the necessary tools for explaining the moral significance of such characterisations. Similarly to framing effects, it is reasonable to expect that, while we could, in principle, explain moral disagreement over any instance of doing/allowing classification, there will be particularly robust and straightforward cases in which appealing to different moral principles would be less justifiable. To see this point, let us take a case which is problematic for purely counterfactual models, like Starving one's baby. In this example, we could argue that general agreement over the classification of this conduct as doing harm reflects the fact that the norm "take care of your child" is much less controversial than the one describing the content and boundaries of a doctor's duties. For this example, the self-contained network model successfully explains why an action which could be defined as omission in some causal models counts for most agents as doing harm.

5.3.4 Summary

In 5.3, I have argued that Hitchcock's self-contained network model can account for three key features of the doing/allowing distinction. Specifically, this model relies on the notion of counterfactual dependence, thus accounting for the fact that doing and allowing amounts to two different ways of causing (i). Second, whether a causal network is self-contained depends on the assignment of default values to all the variables in the causal network. I showed above how different empirical and normative expectations about the "natural course of events" can bring about different value assignments and, in turn, deliver different classifications in terms of doing or allowing. As these different expectations may depend on which description of the case is selected, we may account for the fact that framing affects our doing/allowing attributions (ii). Finally, the self-contained network model can explain why the doing/allowing distinction strikes us as morally relevant; that is, all other things being equal, doing is worse than allowing (iii).

This last desideratum, arguably, requires more elaboration. Specifically, I have not provided here a systematic account of how empirical and normative expectations interact in order to select one value assignment. I give some further indications of this interaction in 5.5. Also, I did not elaborate on which moral features and normative expectations can be incorporated into doing/allowing classifications, and thus imbue this distinction with moral relevance. In Chapter 6, I will argue that those features with obvious and indisputable moral value amount to a) whether an agent had the intention of causing harm and b) whether an agent acted against some salient moral duty or norm.

In the following section, I will provide a final formulation of the alternative thesis, which incorporates the self-contained network account as a tool for making doing/allowing classifications.

5.4 Final formulation

We can conclude this survey of different theoretical proposals, and their relative merits, with the following formulation of the alternative thesis:

> An action amounts to doing harm if the harmful outcome counterfactually depends on the action, within a self-contained network account; an action amounts to allowing harm if the harmful outcome counterfactually depends on the action, within a non-self-contained network account.[17] Whether a causal account is self-contained depends on the value-assignment to variables and, more specifically, on what is set as the "default" value for each variable in the network. An assignment of default value reflects empirical expectations about which course of events is more natural and likely to occur, as well as normative expectations about which moral rules and conventions are appropriate in a

[17] Again, I do not refer here to Hitchcock's further condition for discriminating between real allowings and background conditions.

> given context.

This account tracks the fact that, with doing/allowing considerations, we are interested in the specific way the agent brought about the outcome. Given that different attributions of default value may be reasonable and legitimate in some contexts, the alternative thesis also accounts for the fact that this distinction may be ambiguous, that is, there could be no univocal characterisation in terms of doing or else allowing. As such, this proposal allows for disagreement and framing effects. Finally, while the doing/allowing distinction might not be morally relevant per se, it incorporates normative features of the context, and it is thus morally significant.

5.5 Conclusion

I conclude this section with some final remarks about the rationale of reductionist projects. As I argued above, I define the alternative thesis as an reductionist account, insofar as it argues that the doing/allowing distinction is not a fundamental one, but rather tracks other empirical and normative features of the case at issue. While some reductionist accounts of the doing/allowing distinction amount to influential frameworks, like Bennett's, there are also reasons to be sceptical about the feasibility of the reductionist task. As my discussion in Chapter 2 showed, it is often the case that reducing all doing/allowing classifications to another dichotomy or moral principle fails to account for at least some of the ways we use this distinction in everyday life. In this sense, the distinction seems to be much more nuanced and complex than what a "simple" reductionist model would allow for. Table 5.2 also seems to confirm these concerns: when we strive to reduce the doing/allowing distinction to (i) causal relations, (ii) cognitive features, or (iii) other moral principles, it looks like we inevitably leave some other aspects of this dichotomy out of the picture.

McMahan (1998), in his discussion of Bennett's account, expresses the same concerns about the feasibility of a reductionist project. Despite the merits of Bennett's analysis and its unquestionable clarification results, McMahan is doubtful that this (simple) reductionist account successfully explains the complexities of our use of the doing/allowing distinction:

> "Bennett is undoubtedly right that, in the broad range of cases in which we detect a moral asymmetry between making and allowing, we cannot be confident that we are responding to a single factor that in present and operative in them all. In part this is because, in some of these cases, our intuitions are being aroused and prodded by something quite different (...)

> but another part of the explanation may be that the difference between doing and allowing is not reducible to a single factor (...) It might instead be that, while we are indeed responding to somewhat different factors in different cases, we are also right in detecting an asymmetry between doing and allowing; for the distinction between making and allowing might be internally complex, compounded from various factors that engage our intuitions."[18]

I agree with McMahan that there are reasons to doubt that a reductionist task can be performed in a simple and systematic way. The doing/allowing distinction while clearly tracking *something*, does not seem to simply map onto one single moral principle, but rather captures what is salient for a specific agent on a case-by-case basis. I do not think, therefore, that we can identify a "small" and fixed set of principles or features, or isolate a set of moral principles or categorisations which neatly govern the way we use this distinction in everyday life.

The way the alternative thesis is formulated, however, allows for a more nuanced interpretation of the reductionist rationale. Using the self-contained network account, we can explain on a case-by-case basis which empirical or normative features are incorporated into our doing/allowing classifications. Furthermore, the final doing/allowing attribution can often be better understood as a "summary" or composite judgement of a number of different considerations and expectations, both descriptive and normative. This composite judgement will reflect the fact that we think that the way an agent brought about the outcome matters for doing/allowing attributions, but also that we can have specific moral expectations about how the agent should behave, and these considerations also affect these characterisations.

To be sure, not all moral considerations will be incorporated into doing/allowing classifications at all times. This explains why, in some cases, we may think that allowing harm is morally impermissible (walking past someone who is drowning), or doing harm is permissible (like turning the trolley to save five), seemingly appealing to further rules and principles besides the doing/allowing distinction. Arguably, in a *reflective equilibrium*-style explanation, this might happen when the scenario we are confronted with is resistant to such summary judgements because it pitches against one another different values or considerations in a fairly extreme way. For instance, a behaviour can count, with respect to the agent's causal contribution, (i) as a straightforward instance of allowing, yet is also clearly violating a moral norm (walking past someone who is drowning), or (ii) as a straightforward instance of doing harm, yet it seems the preferable option, if we are forced to choose something (turning the trolley). In such cases, we might not find "a right balance", and thus conclude that the behaviour amounts to doing harm, but it is still permissible, or allowing harm, yet impermissible, and so on, as a way to preserve both intuitions. In other cases, however, when the causal contribution or the norm violation is

[18] McMahan (1998), p. 402.

more nuanced, this summary judgement can deliver a truly “all- things-considered” verdict. In the case of active euthanasia, for instance, agents might balance the fact that the doctor is administering the drug, yet the patient is terminally ill, and the belief that doctors should aim at healing, but also respect patients' wishes. In this example, as we can see from the different positions in the debate, it looks like individuals who think that active euthanasia is permissible would tend to define it as allowing harm, while individuals who think that this is impermissible tend to define it as doing harm. Osman's (2005) and Sinnott-Armstrong, Mallon, McCoy and Hull's (2008) experimental results also indicate that people's moral reasoning seems to reflect this reflective equilibrium approach, and to adjust different principles and values to deliver a “all-things-considered” judgement.[19]

In conclusion, the alternative thesis is characterised as a reductionist proposal insofar as the doing/allowing distinction incorporates other normative and empirical considerations. In this sense, the doing/allowing distinction might not be morally relevant *per se*, as it tracks some causal facts about behaviours, which might not seem of obvious moral relevance, and some separate rules of conduct. Nonetheless, the doing/allowing distinction is still genuinely morally relevant as it delivers these “composite” judgements, which can account for different moral principles or normative expectations depending on the case at issue. Frame-dependency, in this respect, does not undermine moral relevance, but rather explains how different features, whether descriptive, cognitive or normative, could contribute to the final doing/allowing classifications.

In the next chapter, I will elaborate on this proposal, focusing on two main normative features which may contribute to such “composite” judgements: whether an agent intended an outcome, and whether the agent openly violated a moral norm.

[19] Specifically, these experimental surveys show how moral reasoning is “dynamic”, in the sense that individuals tend to adjust their moral judgements as the result of exposure to further information about cases, in a way that strives to balance and retain different intuitions and principles.

6. Intentions, Norms, and Difficult Cases

6.1 Introduction

The alternative thesis argues that the doing/allowing distinction tracks whether an agent is causally relevant to an outcome in a specific way, which is captured by the self-contained network account. Whether the causal network connecting the agent to the outcome is self-contained, however, depends on the value assignments to variables within the model, and, in particular, on which value is set as the default for all variables. These preliminary assignments incorporate agents' expectations and judgements of descriptive and normative features of the context: which course of events is natural, which behaviours are standard, and which actions are morally required. In this way, doing/allowing classifications are based on other moral convictions and the agent's framing of the choice situation; they reflect a "composite" judgement that incorporates separate moral as well as empirical expectations and the fact that an agent is relevant in a certain way to the outcome. If this is correct, the doing/allowing distinction is morally significant as it (also) captures moral considerations, but cannot be ultimately reduced to a single moral principle that consistently explains all doing/allowing attributions. The aim of this chapter is to further support the attractiveness of this thesis.

In sections 6.2 and 6.3, I examine in more detail which morally relevant features can be incorporated into the value assignment to variables. I argue that doing/allowing attributions usually track, in particular, a) agents' intentions and b) which norms and rules of behaviours are salient for the case at issue. In this sense, doing/allowing attributions usually match, in familiar and straightforward examples, a) the fact that an agent intentionally caused harm and b) the fact that an agent was violating standard moral duties. Doing actions thus track intended harm, or harm that violates standard norms, and are suitably harder to justify, whereas allowing actions track unintentional harm, or harm that does not violate any uncontroversial norms. This hypothesis also explains why, in contexts where norms are uncontested and judgements of intentions are clear-cut, agents agree upon doing/allowing characterisations, and their relative moral value. On the other hand, the more cases are under-described, intricate or unfamiliar, in such a way that we are unsure about agents' intentions, or that the normative expectations are contested or controversial, the more agents can reasonably disagree over a) and b), as well as over factual features of the case, thus justifying different doing/allowing attributions. When this is the case, doing/allowing classifications tend to hide the source of moral disagreement, or the fact that agents could "fill in the gaps" in different ways.

In particular, in 6.2 I attempt to analyse the relation between the doing/allowing and the intending/foreseeing distinctions. I show how these two distinctions do not overlap, but they might both refer to the same background expectations about what is "normal" or "conventional". To do so, I describe William Fitzpatrick's (2006) account of intentionality, which uses a so-called "constitutive relation" to discriminate between outcomes that are intended and outcomes that are foreseen. This relation can be further cashed out in natural or conventional terms. I argue that this appeal to "natural" and "conventional" may capture the same expectations about the salient norm/natural course of events that can be used to make doing/allowing considerations. As such, doing/allowing and intending/foreseeing might pick up, in specific conditions, the same features about human agency.

In 6.3, I briefly come back to how norms produce expectations about which course of events should happen, and, in turn, impact doing/allowing classifications. When an agent violated a perceived salient norm, in fact, the associated causal network is often self-contained (that is, the deviant behaviour amounts to a complete explanation of the occurrence of the deviant outcome). I comment on how norm-violation and judgements about intentionality can pick up different expectations, thus delivering different judgements. I conclude discussing disagreement.

The alternative thesis thus matches our strong intuition, in both everyday life and moral theorising, that doing behaviours are harder to justify than allowing behaviours: in clear-cut and transparent cases, doing behaviours consistently map situations where an agent is causally relevant to a harmful outcome, can be described as intending a harm, openly ignores a moral duty or violates a salient rule of behaviour. Allowing behaviours, on the other hand, usually map situations where an agent is causally relevant to a harmful outcome, but the harm does not seem to be intended and the behaviour does not openly violate standard rules.[1] In most examples, arguably, the doing/allowing distinction tracks these aspects so straightforwardly that we do not need to work out which moral principles or normal course of events we are referring to. At the same time, persistence of interpersonal disagreement and evidence of framing effects point to the fact that this categorisation may be ultimately ambiguous in more "difficult" scenarios. When this happens, we should abandon disagreement over doing/allowing classifications and strive to explain and justify our moral judgements at a more fundamental level, and check whether they rely on different empirical and normative expectations.

In Section 6.4, specifically, I use the self-contained network model to discuss some problematic cases, and I show how disagreement can be accounted for within this framework. While I do not conclude that my model accommodates all difficulties, I argue that it provides a useful tool for

[1] More precisely, the allowing behaviour might still be violating a norm, but the agent's norm violation is not perceived as a sufficient explanation for the occurrence of the harm.

explaining disagreement on a case-by-case basis.

6.2 Doing/allowing and intending/foreseeing

The present section, together with 6.3, sets out to examine which morally relevant features can be incorporated into the doing/allowing classification by way of setting the default values of variables. While I have already discussed, in general, how normative considerations can be incorporated in these value assignments, my aim here is to provide a more systematic explanation to this end. Specifically, in this section I will argue that the doing/allowing distinction often matches the fact that an agent intended a harmful outcome, and did not merely foresee it. My contention is that both distinctions, in standard cases, may rely on the same expectations about what is the "normal course" of events, thus delivering the same verdict. Incidentally, this observation also explains why controlling for intentions in fully-equalized cases might prove difficult. First, however, in 6.2.1, I clear the air by showing that the two distinctions do not simply collapse into one another.

6.2.1 Two separate distinctions

In chapter 3, I relied on an intuitive idea of what counts for an agent to intend an outcome, as opposed to merely foreseeing it, for my discussion of *intentions*. In the moral literature, the notion of intention has often been used to explain our intuition that intending a harm is morally worse than foreseeing it, which seems to be at play in this famous pair of cases by Jonathan Bennett:[2]

> **Strategic Bomber**. A bomber drops a bomb on an enemy munitions factory, intending to destroy the factory and thereby damage the enemy's fighting ability, foreseeing that the fallout from the resulting explosion will cause the death of a number of innocent civilians living near the factory, but not intending these deaths.
>
> **Terror Bomber**. A bomber drops a bomb on an enemy munitions factory, intending the resulting explosion-fallout-caused deaths of a number of innocent civilians living near the factory, as a means of terrorizing the rest of the enemy population into giving up the war effort.

[2] Bennett (1980), "Tanner Lectures on Human Values", p. 95. This pair of examples has been further discussed in Bennett (1995), p. 201, as well as in the subsequent literature on the DDE (to cite a couple, McMahan 1993, Bratman 1987). This formulation is from Nelkin and Rickless (2015), p. 378.

Intuitively, the behaviour in Terror Bomber seems harder to justify than the behaviour in Strategic Bomber. This distinction appears to be captured by the fact that, in Strategic Bomber, the bomber merely foresees the death of the civilians, without intending it, while in Terror Bomber the bomber intends the death of the civilians. Here, I attempt to analyse how this distinction relates to doing/allowing, by appealing to a specific account of intentionality, namely William Fitzpatrick's (2006). Before turning to Fitzpatrick's model, I briefly discuss how the two distinctions are in fact separate.

In the discussion of the doing/allowing distinction, the difference between intending and foreseeing has often been perceived as overlapping with the difference between doing and allowing. The concept of intentionality, *prima facie*, could convincingly account for the way we use doing and allowing labels, and, more importantly, explain why doing is worse than allowing. If doing harm mapped cases where individuals intend the harm, and allowing harm mapped cases where an individual merely foresees it, we would have a reasonable justification for our intuition that doing harm is worse than allowing harm to occur, all else being equal. This apparently straightforward explanation seems to work well for some cases, like the pond example. Not aiding a person drowning, which amounts to an allowing-action, might have the same consequence of drowning a person, which amounts to a doing-action, but while in the former instance we merely foresee the harm our action brings about, in the latter we intend this harm directly.[3]

While reducing doing and allowing characterisations to the intending/foreseeing distinction sounds promising, this simple strategy does not successfully account for all cases. As Kagan (1989) notices, indeed, these two dichotomies are not always equivalent. The case of euthanasia, for example, is often employed to illustrate this point: the discrimination between active and passive euthanasia (killing/letting die), which is often regarded as morally relevant, is not captured by the intending/foreseeing dichotomy; in this context, the actions described as “killing” and “letting die”, respectively, are equivalent with respect to intentions. In both cases, indeed, the intentions of the doctor who administers a lethal drug or unplugs a life-sustaining machine seem to be to bring about the death of the patient, as a means of alleviating suffering that cannot otherwise be alleviated. The characterisation of one action as killing and the other as letting die, Kagan concludes, thus refers to a different feature of actions that the intending/foreseeing distinction does not seem to fully capture. Rachels, famously, further

[3] In this discussion, I assume that intentions amount to a morally relevant feature of actions, and thus, all things being equal, intending harm is harder to justify than merely foreseeing it. While this assumption seems intuitive, and has been variously defended (most recently in Nelkin and Rickless (2015) and Victor Tadros (2015)), there are some critical voices, and why and how intentions are morally relevant is contested even among deontologists. Nonetheless, I will set this controversy aside here, and assume that the intending/foreseeing distinction is morally relevant.

argues that we can disentangle the doing/allowing distinction from intentions, using the fully-equalized cases of Smith and Jones. In his example, we have two actions that can be apparently characterised as doing and allowing respectively, despite the fact that we know that both Jones and Smith intended the death of the cousin for the sake of his inheritance.[4] Similarly, in Bystander, where the bystander can choose to turn the trolley, Switch amounts to doing harm to one person; nonetheless, we can reasonably argue that the intention of the agents are to save five people, while she merely foresees the death of the one as an unfortunate side-effect.

In the following section I argue that, while the doing/allowing distinction does not simply overlap with the intending/foreseeing distinction, in many cases our judgements about intentionality rely on the *same* expectations about which actions and outcomes are appropriate or more likely to happen in the given situation. By way of setting the default value of variables, considerations about intentionality therefore match the classification of actions as doings or allowings in such cases. When this happens, we might conclude that the two distinction capture *the same feature of human behaviour*.

6.2.2 Intentions and expectations

To justify this position, we need to examine more precisely to the concept of "intention" and to how it has been theorised in moral philosophy. I do not attempt here a thorough survey of accounts of intentionality; to navigate my way through the literature, I will draw on a problem for all moral principles that focus on intending harm.[5] This difficulty was first raised by Jonathan Bennett (1995, 194–225); following Nelkin and Rickless (2015) and others in the literature, I call this the problem of *closeness*.

6.2.2.1 The closeness problem and Bennett's solutions

The problem, roughly, amounts to the fact that, for any instance of "intending", it is possible to re-describe the situation so that the agent did not intend the harm, but rather "intended some other state of affairs that is causally responsible for (or otherwise connected by some relation

[4] Rachels, specifically, argues that the example shows that there is no moral difference between the two cases, and therefore between actions characterised as doings and actions characterised as allowings. Nonetheless, this pair of cases pulls apart the doing/allowing and intending/foreseeing distinctions, whether or not we think the former is morally significant.

[5] See, among others, Foot (1978).

other than identity to) harm".[6] To illustrate this point, Nelkin and Rickless design the following example:

> **Sophisticated Terror Bomber.** A bomber drops a bomb on an enemy munitions factory, intending only that civilians' bodies should be in a state that would cause a general belief that they were dead, this lasting long enough to shorten the war: nothing in that scheme requires that the dismaying condition of the bodies be permanent; so nothing in it requires that the civilians become downright dead rather than merely seemingly dead for a year or two.[7]

In some sense, the Sophisticated Terror Bomber can amount to a more "fine-grained" description of Terror Bomber, where the agent cannot be described as intending the death of the civilians. In the active/passive euthanasia example, we can similarly give a more fine-grained description of the actions of the doctor who unplugs the machine, arguing the she did not intend to harm or kill the patient, but merely to stop the pain by stopping the patient's brain from working, foreseeing that this would kill the patient. In Bystander, we seem to already be using a fine-grained description in order to identify the intentions in Switch as "saving five" rather than "killing one". But which of these re-descriptions is reasonable or legitimate, if we want to account for our intuitions that intending harm is worse than merely foreseeing it? To be sure, not all these ways of explaining an agent's behaviour could be acceptable if we want to conclude that, at least in some cases, an agent did intend to harm. In short: arguably, a harm hardly ever seems to be the final or "desired" goal. Hence we can always re-describe a case, arguing that only the final goal was intended, and the harm only foreseen. Clearly, however, as Bennett argues, we also need to "draw a line": sometimes this re-description is legitimate, but at other times it is not.

This question is by no means an easy one to answer. Bennett (1995, pp. 205–210), for instance, attempts different strategies to solve the closeness problem. One involves checking whether two descriptions are actually descriptions of the *same event*. If this is the case, we cannot reasonably argue that the agent intended the harm in the one description but did not intend it in the other.[8] Note that, however, if we concede that events are individuated at least partially by time of occurrence, then, in the euthanasia example, "stopping the pain" or even "administering a lethal

[6] Nelkin and Rickless (2015), p. 380.

[7] Nelkin and Rickless (2015), p. 380. The formulation of this case draws on Bennett's (1995) discussion of similar examples (p. 202–208). Bennett (p. 204) also illustrates the problem of closeness using the following *child* case:
"A nearly born child is blocked; its mother is near to death, and her heart cannot stand a Caesarian delivery; to extract the child, the surgeon crushes its head, thereby killing it. Did the child's dying lie within the scope of what the surgeon intended, or did he intend only to change the shape of its head, its death being a foreseen but unintended by-product?"

[8] Bennett (1995), p. 206.

drug" can legitimately be considered as separate events with respect to "killing the patient", as they do not occur at the same time.[9] This conclusion seems to be inconsistent with our intuitive judgement that, by administering the drug, the doctor intended the death of the patient, and that the re-description is not "legitimate".

Other proposals draw on the concept of causal necessity. As Bennett further argues (p. 209), maybe it is the case that two states of affairs amount to two separate events. But if one state of affairs "causally necessitates the other (in the sense that the first makes the second inevitable), it is impossible to intend the first without intending the second".[10] This proposal, nonetheless, also seems to fail in capturing our intuitions about cases. In Bystander, for instance, the way the example is described, it is contingently impossible to save five people without inevitably killing one. The bystander cannot thus be described as merely foreseeing the death of the one person; this arguably challenges our common and more natural intuitions.[11] After exploring and abandoning other possible proposals, Bennett concludes his investigation by arguing for the following "loose" solution to the closeness problem:

> "The best I can find is rather loose, but it may be the whole truth about our intended/foreseen distinction. Not only is there no chance of (…) crushing the baby's head without killing it [or to bombing the civilians without killing them] – these things are what the plain man would call *inconceivable*."[12]

6.2.2.2 Fitzpatrick's constitutive relation and the appeal to background expectations

William Fitzpatrick's (2006) proposal draws on, but strives to elucidate, this concept of "inconceivability". Fitzpatrick argues that:

> "if the relation between two states of affairs is known to the agent, natural and constitutive rather than merely causal, then we cannot properly speak of an agent's intending the one while merely foreseeing but not intending the other".[13]

I focus here on Fitzpatrick's account of this *constitutive relation*. According to Fitzpatrick, this

[9] Or, as Bennett puts it with respect to the *child* case (footnote 7), "the collapse of the head and the death of the [fetus] (...) occur a second apart".

[10] Nelkin and Rickless (2015), p. 383.

[11] In the child case, "when a hysterectomy is performed early in pregnancy, (...) it is causally inevitable that the [fetus] will die." (Bennett, p. 209).

[12] Ibidem, p. 213.

[13] Fitzpatrick (2006), p. 603.

relation is different from identity, logical entailment or causal necessity, and defines a connection between states of affairs that can be *naturally* or *conventionally* determined. A constitutive relation is naturally determined when it is "simply a matter of natural fact"[14]: for instance, he argues, being blown to bits constitutes being killed; nonetheless, being blown to bits is not equal to being killed– neither does it logically entail being killed. Also, a constitutive relation can be "determined by merely conventional arrangements involving the agency of others".[15] An illustration of this is the following example: a school's conventions and rules stipulate that if a student fails an exam she is kicked out of the programme; failing an exam thus constitutes being kicked out. Fitzpatrick's solution seems to account for our intuitions in most familiar examples: administering a lethal drug naturally constitutes killing a man, while turning the trolley does not constitute killing a man, neither naturally nor conventionally; this also matches our common attributions of intentions in the two cases.[16] This account, I suggest, is onto something promising, and can serve us in "drawing a line" between harms that are intended and harms that can be described as foreseen: an agent can be said to intend harm if the state of affairs that is a result of her actions constitutes harm.

This framework, besides providing a suitable solution to the problem of closeness, also helps us establish a connection between intentions and doing/allowing classifications by appealing to *background expectations.* Note that the constitutive relation relies on what is natural, common and expected in a given situation, so as to determine whether a state of affairs conventionally constitutes another. The self-contained network model, which I use here as a tool for tracking doing/allowing classifications, similarly relies on the prior definition of what is "standard" or "normal" within a specific context. These background expectations that determine judgements of intentions could thus plausibly amount to the same empirical and normative features that "set the scene" in the self-contained network model by way of setting the default value of variables. I then suggest that the doing/allowing distinction and the intending/foreseeing distinction overlap exactly in those cases where they rely on the *same* judgements about what is the "normal course of events". In these situations, the two distinctions track the same morally relevant fact that intending a harm is worse than not intending it.

I now discuss this hypothesis in more detail. Fitzpatrick's account of intentionality argues that, for an agent to intend an outcome S2, it is not sufficient that she intended S1, and that S1

[14] Ibidem.

[15] Ibidem.

[16] Nelkin and Rickless (2015, p. 391) argue that this solution faces some difficulties as well, especially because the constitutive relation is never defined by Fitzpatrick but only illustrated with the help of some paradigmatic cases. Nonetheless, I think that this account captures some important intuitions, and thus assume that problems with this framework could be fixed, by way of defining the constitutive relation more systematically.

constitutes S2; the agent also has to *know* that S1 constitutes S2.[17] I am not specifically interested here in how Fitzpatrick uses this knowledge condition, but rather in the underlying intuition: to make judgements about intentionality, we have to know something about the agent's mental states, beliefs and desires. Maybe the professor does not know that if she fails the student she will be kicked out, and she merely wants to be fair; maybe the doctor does not know that injecting a lethal drug will inevitably kill the patient. In these cases, we might be ready to concede that the professor and the doctor did not intend to kick out the student and kill the patient, respectively. What happens, however, when we do not have direct access to the agent's mental states? Arguably, to infer agents' intentions, we can only look at actions, and how these actions are related to the outcome. If what an action brings about naturally or conventionally constitutes S2, we might thus conclude that the agent intended S2. On the other hand, if the outcome of an action does not naturally, usually or conventionally constitute S2, we might infer that the agent did not intend S2.

I suggest that we can cash out Fitzpatrick's definition of the constitutive relation, generalizing it for cases where we lack direct access to an agent's mental states, as follows:

> X constitutes Y if, absent other information, it is widely recognized that X is one of the ways in which Y can be reliably realized.

By way of elaboration: if an action is the most natural, standard or conventional way to realize an outcome, we might conclude that the agent intended such an outcome, especially if we have no access to the agent's mental states. Injecting a lethal drug, arguably, amounts to one of the most natural ways of bringing about death, and thus we will conclude that the doctor intended to kill the patient, relying on what we think most doctors would and should know in this context. The case of the professor, as Fitzpatrick concedes, is much trickier: if we judge that the convention is strong enough, and that most professors who want to kick a student out usually fail them, we will conclude that the professor intended to kick the student out. On the other hand, we could also reach a different conclusion, depending on how closely associated we think failing is with being kicked out. To be sure, these inferences about intentionality only amount to our “best guess” in this respect, absent more detailed information about the agent's motives in performing an action.

[17] Fitzpatrick (2006), p. 595.

6.2.2.3 Intentions and causal networks

Let's now see how these thoughts translate in the assignment of default values in the self-contained network model, and, in some practical cases, how doing/allowing attributions and judgements about intentionality can track the same expectations about what is normal or natural.

In the pond examples, for instance, the intending/foreseeing distinction and the doing/allowing distinction overlap. Within the self-contained network model, for drowning the person to count as doing, we would need the outcome “death” to take a default value when its parent “push the person under the water” takes its default value as well. For the failure to rescue to count as allowing, on the other hand, we would need the outcome “death” to take its deviant value while its parent “do not rescue” takes its default one. In Drowning we have:

D = death of the person by drowning, where D = 1 if the person drowns and D = 0 if the person does not drown, and Def(D) = 0.

This attribution can be justified by referring to what is “normal” or expected in a natural or statistical sense. As for O's parent,

P= pushing someone underwater, where P = 1 if I push the person underwater and P = 0 if I do not push the person, and Def(P) = 0.

Clearly, pushing someone underwater amounts to “deviant” conduct, and this explains why the default is set at 0. The causal network {P, D} is self-contained, as the outcome takes its default value when P does. The evaluation of the agent's intentions in this case, arguably, relies on the same judgements and expectations. With no direct access to the agent's mental states, all we can observe is that the agent performs a “deviant” action that brings about the deviant outcome “the victim drowned”. On Fitzpatrick's definition, being drowned naturally constitutes being killed; thus we recognize that drowning a man amounts to a natural and fairly efficient way to kill someone: hence the inference that the agent intended to kill the man. This conclusion matches the classification of this action as “doing harm”.

In the failure to rescue case, we have the same value assignment for D, while its parent J = continue jogging takes the value 1 if I continue jogging and 0 if I stop jogging and rescue the person, and Def(J) = 1. Here, again, the assignment of the default values can be justified by referring to what is normal or standard. In the self-contained network model, the action is classified as allowing, since the outcome can take its deviant value when its parent takes its default one. When inferring the agent's intentions in this second case, what we will see is an action, jogging, and a resulting state of affairs, which is that the agent continues her workout.

This state of affairs does not constitute killing a man, neither naturally nor conventionally; in most situations, indeed, these two outcomes are not even related. Usually, working out does not amount to the most natural or efficient way to kill someone. This explains why the agent might be described as merely foreseeing the outcome. Again, this conclusion matches here the “allowing harm” classification.

In many familiar and clear-cut scenarios, arguably, the self-contained network model delivers classifications of doing and allowing that are often consistent with the intuition that an agent does harm if she intentionally performs an action that results in a harmful upshot. When we are asked to evaluate another agent's behaviour, and we do not have direct access to her motives and desires, doing/allowing classifications are thus usually a reliable proxy for tracking our *judgement* of whether the agent acted intentionally, which I assume amounts to a morally relevant feature of actions. I suggest that, in the case under scrutiny, both distinctions rely in fact on a prior judgement about which course of events is more natural or normal. In many simple scenarios, doing harm describes a situation where an agent performs a specific action that usually brings about the harmful upshot, and in most cases does so by “deviating” from the normal course of nature. In these examples, we can reasonably and legitimately infer that the agent had the intention to harm. Many allowing cases, on the other hand, amount to situations where the agent performs an action that does not bring about a state of affairs that constitutes harm: in most failure to rescue cases, for instance, the outcome of the action is simply the continuation of one's activities, and does not deviate from the standard course of events.[18] In those situations, our best guess – absent direct access – is that the agent did not intend to harm. In such contexts, the self-contained network model can, in conclusion, explain the intimate connection between intending and doing, if we accept Fitzpatrick's analysis of what it means to intend harm. Note that this also explains the difficulty in controlling for intentions when comparing fully-equalized cases.

To be sure, there will be cases where these two distinctions do not match, and it is thus possible, for instance, to do harm without intending it, as in Bystander,[19] or to allow harm while intending

[18] Note that doing behaviours will usually equate to the variable taking its *deviant* value. It might thus seem counterintuitive that, for an action to count as “doing”, it has to be the case that the outcome takes its *default* value when its parent takes its default one. This condition, however, determines whether the causal network connecting the agent to the upshot is self-contained, i.e., the explanation for the outcome occurring strikes us as “complete” without referring to other “parents”. In this sense, it is reasonable to think that I “do harm” when I perform a deviant action, and this is sufficient to explain why the outcome takes its deviant value. On the other hand, for my action to count as allowing, it has to be the case that the outcome can take its default value when its parent takes its deviant one. One possible configuration of allowing behaviours is thus that I perform a default action, and the outcome is deviant (and therefore my action does not look like a sufficient explanation for the outcome occurring).

[19] At least according to Thomson's interpretation.

it, as in Jones's example. I will discuss these two cases in detail in section 6.4. For the moment, however, note that doing/allowing classifications might not always track intentions simply because intentions are not all that matters in justifying and explaining the doing/allowing distinction. Firstly, as discussed in 6.2.1, intending harm is not equal to doing harm, and thus the doing/allowing distinction also captures a descriptive fact about the agent's causal contribution to an outcome: whether the outcome occurred or not and how the agent contributed to the outcome. Secondly, the explanation above mostly works for cases where we have to infer intentions from actions, and thus our judgements about intentionality rest on background expectations and draw on similar cases and presuppositions about what is “normal”. These judgements, however, are not particularly robust, and can be reversed if we gain additional information about the agents' motives and beliefs. Moreover, doing/allowing classifications might track other normative considerations as well, which can also determine the value assignment to variables. Specifically, I argue that a given variable might take a deviant value if it openly violates a rule of conduct or salient norm. In 6.3, I move on to consider how doing/allowing classification can be sensitive to norm-violation as well.

6.3 Salient Norms

In this section I examine in more detail the connection between the doing/allowing distinction and norm-violation. I argue that doing/allowing classifications may also track how an agent's behaviour fits in with moral and social principles, rules and expectations. The self-contained network model allows us to explain how these considerations may ultimately determine whether an action is perceived as an instance of doing or allowing by way of setting the default value of variables. As norm-violation arguably amounts to a morally relevant feature, my account thus shows that the doing/allowing distinction is morally relevant. At the same time, as different individuals can perceive different norms as salient, or disagree about the import of different moral principles, this model also accounts for the fact that doing/allowing classification might be ambiguous, i.e., agent-, context- or frame-dependent.

My general hypothesis is that a norm creates an expectation that people should act on it. This expectation, in turn, determines the default value attached to the variables. Usually, if an agent's behaviour violates a norm and a harmful outcome occurs, the causal network connecting the agent and the outcome will be self-contained, as norm-violation (deviant) is perceived as a sufficient explanation for the variable taking its deviant value. On the other hand, the fact that a harmful upshot counterfactually depended on the agent, but the agent did not seemingly violate a norm, often results in a non-self-contained causal network: the action (default) is not perceived as a complete explanation for the occurrence of the harm (deviant).[20]

In section 5.3.2, I discussed the case of active euthanasia along these lines. Another interesting example of “doing as norm violation” could be Starving one's baby. Here, upon counterfactual analysis, “not feeding” could amount to allowing, thus sharing the same causal structure as those instances Hitchcock identifies as omissions (see Figure 2.3). Nonetheless, the interpretation that I “did harm” my baby seems reasonable and tenable as well: in this specific context, the fact that parents should care for their children amounts to a particularly stringent social norm. As such, we might expect that the parents should feed their baby, and that a baby should not “naturally die”, like the flowers my neighbour forgets to water. When we set these different default values, the parents' behaviour takes the “doing” classification.

I do not offer here a more systematic account of which specific norms can be incorporated into doing/allowing classifications. My contention is that almost any morally significant expectation could do the job. Arguably, the more a moral norm is agreed-upon or stringent, the more likely it is that our final doing/allowing classifications will track this norm. Also, the less “morally sensitive” a case is, the more even trivial expectations about standard behaviours could inform the value assignments to variables; these cases, however, are also likely to be less agreed upon with respect to doing/allowing characterisations. In short, when it comes to setting the default value of actions, we refer to what the “normal” course of events is, and determining what counts as “normal” comes down to our empirical expectations about what will or should happen. The presence of strong moral or other social norms can mean that our empirical predictions or expectations are more precise and agreed upon.

In the remainder of this section, I briefly discuss how norm-violation can be based on different background expectations with respect to those used to make judgements about intentions (6.3.2), and what the appeal to norms can tell us about disagreement over doing/allowing classifications (6.3.3).

[20] In Chapter 2, I suggested that allowing harm might be characterised as well by norm violation, but within a non-self-contained causal network. These two possibilities amount to two different configurations of “allowing harm”: the behaviour takes its default value and the outcome takes its deviant one; or the behaviour takes its deviant value and the outcome takes its default one. The first case is the one I am focusing on here: a harm occurs, which is causally dependent on the agent, but the agent is not perceived as violating a norm. The second case captures situations where the agent might be violating a norm, but this norm violation on the agent's part is not a “sufficient” explanation for the harm occurring. My intuition is that the more a norm is perceived as stringent or “important”, the more its violation would be perceived as a sufficient explanation for the harm, thus setting different default values to variables.

6.3.1 Norms and intentions

Concluding my discussion of intentions, I argued that, in some cases, judgements of intentionality and the doing/allowing distinction pull apart. In most examples, the easiest explanation for this fact is that we get more information about an agent's desires and beliefs, and thus we do not need to "guess" her intentions: if I step on your foot, and I say "sorry– I did not mean to!", I obviously "did harm", but the harm is also unintended. Some more nuanced examples could be cases where an agent so openly violates a specific social norm or rule of conduct that her action is characterised as doing harm, even if the harm was unintended. One possible illustration is, again, Starving one's baby. In this example, relying on Fitzpatrick's analysis above, we might argue that "not feeding a baby" does not *constitute* the death of the baby. Consistently, the action of starving one's baby can be described as not intentional, yet "doing harm".

Note that this interpretation seems to be consistent with the idea that, for someone to count as doing harm, we do not necessarily require that one is motivated by particularly "evil" intentions. While the fact that intentions are not all that matters in our comprehensive moral judgements is not surprising, this observation has some particularly interesting implications for the case of bringing about harm. As Philip Pettit (2015) notices, doing good seems to require us to fulfil some specific and fairly demanding set of norms and expectations; doing harm, however, is not symmetrical with doing good in this respect. For someone to count as doing harm, indeed, we do not require her to meet specific standards or rules, but it is often sufficient to show that she failed to comply with what was asked of her to "be good". In this sense, Pettit observes, an agent often counts as doing good, or helping, only if she is doing good "for good's sake", or if she is specifically fulfilling a moral duty.[21] On the other hand, for an agent to count as doing harm, we do not require her to be doing evil "for evil's sake", or to have a particularly callous motivation: it can simply be the case that, for her own interest, laziness, or other trivial reasons, she is not complying with what it is expected of her. Relying on the model of intentionality discussed in the previous section, it can thus be the case that an agent does not intend to harm, because her actions, as in Starving one's baby, bring about a state of affairs that does not constitute harm. On the other hand, the agent might not be complying with what is required of her, and this could be sufficient reason for her "doing harm".

[21] Pettit (2015), pp. 172–195.

6.3.2 Norms and Disagreement

In this final section, I explore the fact that different individuals might disagree about doing/allowing classifications because of normative features of the context. When this is the case, different classifications track different expectations about what *should* happen. I suggest that disagreement here is most promisingly investigated at the more fundamental level of normative expectations.

Note that disagreement about which default values are appropriate can have two main sources. First, different individuals can perceive different norms (or no norms at all) as salient. This phenomenon could depend on a variety of factors: cognitive biases, idiosyncrasies, past experiences, or social or cultural background. The fact that one specific rule of behaviour rather than another is used to set our expectations in a given scenario, at the end of the day, appears to be ultimately an empirical matter, which should be investigated accordingly. Lacking the empirical data in favour of or against any of the above hypotheses, I do not favour one or the other. I suggest that, following my discussion of framing effects, there are reasons to suspect that people would converge on the salience of norms when cases are more familiar, straightforward and detailed. Secondly, disagreement might also be substantive: different individuals can have different moral convictions, and this ultimately determines whether they think a specific rule should be followed, a specific obligation fulfilled and so on. Agents might also agree that a set of principles or rules is relevant, but "balance" them differently. In this second case, different doing/allowing classifications may incorporate and reflect different ethical perspectives.

Whatever the source of disagreement, focusing on the doing/allowing distinction might not advance the discussion. Moreover, we should also be more circumspect in using these classifications to make comparative moral judgements. To be sure, these classifications capture something morally significant, but investigating the underlying norms and values could be a more promising way of dealing with disagreement.

In the final section of this chapter, I attempt such an analysis of some of these difficult cases.

6.4 Difficult cases and the alternative thesis

Throughout this work, I have often referred to controversial examples, where our descriptions of doing and allowing are not consistent intrapersonally – i.e., they are frame-dependent – or interpersonally–there is disagreement about the correct doing/allowing classification. Moreover, in some instances, there is also disagreement about whether a doing action is morally worse, harder to justify, or different from an allowing action, all other things being seemingly equal. My account of the doing/allowing distinction provides a specific explanation for this persistent "instability": given that doing and allowing classifications rely on expectations about what counts as the standard course of events, agents may reasonably disagree about what the "normal" state of affairs is. Disagreement can involve both descriptive features of the context and normative aspects, such as which moral norm is more relevant for evaluating the case at issue.

Nonetheless, in familiar and straightforward cases, we will likely observe more stability and agreement in expectations about what will or should happen. As a result, these cases will be less controversial in terms of doing and allowing classifications, and this categorisation will easily match other morally relevant considerations, such as intentionality or norm violations. On the other hand, when agents are asked to evaluate examples which are unfamiliar, extreme or artificial and, in general, far from the ordinary, there will be substantial disagreement over which course of events is to be expected. Difficult cases can also amount to scenarios that are significantly under-described: in these situations, different agents can pick up different cues and "fill in the gaps" in different ways. Leaving aside the case of moral disagreement, arguably, the "difficulty" of cases could be a matter of degree: the more unfamiliar or under-described a case is, the more room there will be for alternative interpretations and reconstructions, and the more tenuous our intuitions about doing and allowing classifications will become.

I examine here some paradigmatic examples, showing how these different explanations for disagreement and instability can be appropriate, on a case-by-case basis.

6.4.1 The Smith/Jones case

When discussing fully-equalized cases, I argued that Jones's story strikes us as "weird" and artificial. I think that now we have more elements to explain the difficulties and persistent disagreement in evaluating this example. The Smith/Jones case seems particularly problematic with respect to Fitzpatrick's account of when an agent intends an outcome. According to the framework outlined above, an agent cannot merely foresee and not intend harm if the outcome that is a result of her action constitutes harm. When we have no access to an individual's mental states and beliefs, I further argued that we will generally infer her intentions from the fact that the actions she performs naturally or conventionally bring about a harmful upshot.

When examining Smith's conduct, this standard way of reconstructing intentionality matches the background story of Smith's desires and beliefs. In this first case, the attribution of evil intention to Smith also tracks the fact that the upshot takes its default value (i.e., the cousin does not die) only when the action it counterfactually depends on does (i.e., Smith does not drown the cousin), and thus the action is classified as doing. What about Jones's conduct? Here, the action performed by Jones is classified as allowing, since the upshot takes its default value when the action takes its deviant one (i.e., Jones saves the cousin, while the "default" would be doing nothing). Note that, if we had no access to Jones's mental states, and we would have to infer his intentions from his behaviour, we might conclude that he did not intend to kill his cousin: after all, the outcome of the behaviour "do nothing" does not constitute, neither naturally nor conventionally, someone's death by drowning. In this second case, however, this reconstruction does not match the background story and the information we have about Jones's mental states. Of course, when we have such "direct" information, we do not need to infer intentions from actions, and therefore we can simply conclude that Jones intended to kill his cousin.

This discrepancy is, arguably, due to the fact that there is a particularly lucky coincidence here– for Jones at least: the cousin is already dying. This "twist" in the story makes Jones's case quite different from most familiar and straightforward examples. Therefore, judgements about intentionality do not match doing/allowing classifications, as the former do not rely here on the same expectations of what is "standard" or "normal", but rather on the more direct information we draw from the background story. Agents might then disagree about whether Smith's conduct is worse than Jones's, because both the doing and the allowing actions are intended. When asked to make a moral evaluation, we could thus appeal to two alternative features that often come together in familiar cases: whether someone intended to harm and whether she contributed to the outcome in a specific way, which is captured here by the doing/allowing distinction.

In conclusion, I do not think that the Smith/Jones case is problematic for doing/allowing classifications, or that it undermines the moral relevance of the distinction. This example rather shows that allowing behaviours do not necessarily amount to cases where the harm is not intended, even if, when we do not have access to agents' mental states and background stories, this is usually a good approximation.

6.4.2 Trolley cases

Trolley cases amount to another class of examples used in the moral literature to argue in favour of or against the moral significance of the doing/allowing distinction. Specifically, at least in

Thomson's interpretation, people seem to disagree on whether doing is worse than allowing. I claim here that these cases are not problematic for my account, which can also explain why these scenarios seem particularly difficult.

Recall Bystander: a bystander has to choose between letting the trolley run over five innocent people (Don't Switch) or turn the trolley so that it runs over over one innocent person (Switch). Thomson defines Switch as doing harm, and Don't Switch as allowing harm. The self-contained network model accounts for this classification: the most natural course of events is let to the trolley continue on its track; the default values assigned to the variables, however, reflect the common expectations that five people are not run over and the bystander does not interfere by switching. The causal network describing this scenario is non-self-contained: it is possible for the outcome to take its deviant value (five people die), while its parent takes its default one (the bystander does not intervene by switching). Don't Switch, therefore, is characterised as allowing harm, since the causal network is non-self-contained. On the other hand, the default value of diverting the trolley is intuitively set at 0 (i.e., the action is not performed), and the death of the person on the other track takes its default value at 0 as well (i.e., this person does not die). The causal network relating the death of one person to Switch are thus self-contained, and this conduct is characterised as doing harm.

To be clear, I think that the self-contained network would also account for a classification of Switch as allowing harm: intuitively, throwing the switch does not seem a sufficient explanation for the death one one person. Therefore, we can work out value assignments to variables which will capture this insight. In what follows, however, I assume that Thomson's analysis is correct, and challenge her claim that these cases are problematic for the moral significance of the doing/allowing distinction. My discussion here mostly relies on my observations in 3.3.2, and just supplements these observations with the conclusion that the self-contained network model can account for different doing/allowing classifications, and it is not undermined by Thomson's argument.

Recall that Bystander, according to Thomson, will elicit different judgements of permissibility of the doing harm conduct with respect to Fat Man. Specifically, while it is impermissible to Push in Fat Man, it is permissible to Switch in Bystander, and this would undermine the claim that people appeal to the doing/allowing distinction when making moral judgements. My objection to Thomson is that doing/allowing distinctions do not necessarily deliver an "all-things-considered" judgement, which balances all morally relevant features of behaviours. In this sense, it is possible for an action to be characterised as doing harm, while at the same time other considerations make this conduct look permissible.

As I argued before, one way to explain the difference between Switch and Push is to appeal to the fact that it is more difficult to argue, in Fat Man, that the bystander did not intend to kill the man when she pushed him onto the track. Being pushed in front of a trolley in motion, indeed, naturally constitutes being killed, and it is what we should naturally expect when performing such an action. We could conclude, following Fitzpatrick's analysis of the closeness problem, that it is not possible for an agent to intend to stop the trolley without intending to kill the man. In this case, therefore, the fact that the action amounts to doing harm *and* can be described as intentional tips the scale in favour of moral impermissibility.

To be fair, different accounts of intentionality and solutions to the closeness problem would allow us to characterise Push in Fat Man as merely foreseeing the death of the man. Other interpretations of what makes Switch and Push morally different have thus been attempted in the literature. Whatever the case, I do not think that trolley problems threatens my account of the doing/allowing distinction and, more generally, any account that claims that this distinction is morally significant. The mere fact that a doing action (Switch) can be considered permissible, or the fact that one doing action (Push) is worse than another doing action (Switch) does not disprove the more specific claim that, all other things being equal, doing is morally worse than allowing. While the comparison between Push and Switch might seem problematic, we could argue that pushing the bystander is worse than turning the trolley because of some further distinction the doing/allowing distinction alone does not capture. For this reason, when we compare these two doing actions with the chance of saving five people, the fact that the former is worse than the latter might influence our final judgements about permissibility

Arguably, trolley cases are particularly unfamiliar, difficult and extreme, and none of the actions described seem to be clearly morally permissible or morally required. It is not surprising, therefore, that people lack strong intuitions about the degree of moral wrongness of these behaviours. Nonetheless, trolley cases help clarify that morality is a very complex matter, and the doing/allowing distinction is only part of a much more comprehensive picture.

6.4.3 Preemptions

In chapters 1 and 2, I discussed some problematic cases for all difference-making accounts of the doing/allowing distinction, which are referred to in the literature as instances of “preemption”. The simplest example of this class is the Backup case, where we apparently cannot classify Alice's behaviour as doing harm, since the outcome would have occurred anyway due to Backup. Settling for the self-contained network account, I argued that we just had to bite the bullet, and allow preemption to be an exception in an otherwise convincing explanatory account. A partial answer to this difficulty is that preemption cases do not seemingly amount to familiar and common situations, but rather to the class of "cleverly devised" scenarios,[22] which rely on the presence of weird coincidences that are extremely unlikely to verify in ordinary circumstances. Nonetheless, the self-contained network model does not successfully account for our intuition that Alice is doing harm, unless we use some further refinements such as the "freezing" strategy.

6.4.4 Borderline cases

In this section, I discuss what I call "borderline cases", i.e., particularly controversial cases where people's intuitions, and positions in the moral literature, disagree on the classification of an action as doing or allowing. I do not survey all these cases here, but I focus on a few paradigmatic examples.

Let's first take this case from McMahan (1993):[23]

> Impoverished Village: Having given one's accountant full power of attorney one learns that because of a misunderstanding he is preparing to sign away 10% of one's income to save the lives of people in a remote impoverished village. One phones to instruct him not to do it.

This example from Rickless (2011) also seems to involve something like "withdrawing aid":[24]

> Hospital: A doctor has just plugged one person into a respirator. If the patient is moved or unplugged from the respirator, he will die. Five more patients arrive and will die unless plugged into the respirator. The doctor unplugs the first patient in order to save the five.

The two following examples are from Barry and Øverland (2017):[25]

> Interpose: A cart filled with water is rolling downhill. Bill, who is sitting at the bottom of the hill, will survive if the cart reaches him. Sue interposes a rock; the cart stops and Bill dies of thirst.
>
> Remove: A cart is rolling towards a point where there is a rock that would bring it to a halt. Sue removes the rock; the cart rolls down the hill and injures Bill, who is sitting

[22] Sunstein (2003), p. 10.

[23] I take this specific formulation of the Impoverished Village example from Woollard (2015), p. 9.

[24] Rickless (2011), p. 68.

[25] Barry and Øverland (2017), p. 85.

> there.[26]

Arguably, most people agree with the classification of Impoverished Village as an instance of allowing harm. The consequently listed examples, however, seemingly describe (increasingly) more morally wrong behaviours, which are nevertheless not clear-cut instances of doing harm. In short, doing and allowing classifications require us to make a sharp distinction between these actions; upon careful analysis, however, these scenarios only differ in small details or nuances. We are thus in the position, wherever we decide to draw the line, to justify the fact that almost identical actions belong to two separate categories. The four examples above, indeed, all describe a situation where the agent, at least according to our immediate reactions, *does* something, as she seems to perform an action or to intervene, and does not merely observe the "course of nature" (like, on the contrary, Jones's behaviour in the pair of fully-equalized cases).

With respect to borderline cases, different proposals in the moral literature draw different lines between doings and allowings. Quinn and Bennett, for instance, could argue that Remove and Interpose are instances of doing, as in both cases the upshot occurs because the agent did something. Our different intuitions about Impoverished Village and Hospital– which are seemingly less morally objectionable– could be explained, within Bennett's account, by pointing to the fact that the outcome would not have occurred in the first place without the agent (as the sum of money would not have been given to charity or the respirator plugged into the first patient). Foot, on the contrary, argues that we can discriminate between these cases by relying on the difference between initiating, sustaining, enabling and forbearing-to-prevent a harmful sequence: while initiating and sustaining would describe cases of doing harm, enabling and forbearing-to-prevent would be cases of allowing. All four cases described above, within this framework, could be instances of allowing, as a harmful sequence is seemingly not initiated or sustained by the agent. In Impoverished Village and Hospital, indeed, the agent stops an aiding sequence she herself initiated, while in Remove the harmful sequence is already in place, and in Interpose the agent stops an aiding sequence.[27]

More recently, Woollard (2015) has argued that:

> "An agent counts as doing harm if and only if some fact about the agent's behaviour is part of the sequence leading to harm; the agent counts as merely allowing harm if and only if a fact about the agent's behaviour is relevant to, but not part of, this harmful

[26] Barry and Øverland (2017) define both Remove and Interpose as instances of *enabling harm*, which they conceive as an intermediate category between doing and allowing.

[27] I do not argue, in this section, that these classifications in terms of doing or allowing are the ones the authors I refer to would give for these specific cases. My aim is rather to show how different accounts can deliver different classifications of these borderline cases.

> sequence."[28]

Recall that Woollard further distinguishes between *substantial* and *non-substantial* facts: the former are intuitively perceived as more "natural" parts of a sequence, while the latter usually amount to background conditions. Specifically, a fact counts as substantial if it is either positive or contradicts normal presuppositions. She then claims that, if an agent is relevant to a harm through a non-substantial fact about her body, then her actions are merely a condition for, rather than part of, the harmful sequence, and this would count as allowing harm. On the other hand, if there is a complete sequence of substantial facts leading from the agent to a harmful effect, the agent's action would count as doing. Let's see how this account could classify the examples above. In Impoverished Village, the fact that the agent does not sign her money away seemingly amounts to a non-substantial fact, so we can classify her action as allowing. These cases, however, could be also explained as "removing barriers" examples, which Woollard discusses as a separate category. In the Hospital case, the respirator might be perceived as a barrier, which does not belong to the patient, nor to the doctor specifically. In some sense, however, if we think at the doctor as a part of the hospital, then the respirator/barrier "requires the continued use of resources belonging to the agent".[29] When this is the case, following Woollard, we could conclude that "the removal of the barrier counts as merely allowing harm. The agent simply refuses to let his or her resources be used to protect the victim."[30] I note, however, that in Hospital the removal of the respirator could also be perceived as a positive fact, or a fact which contradicts our normal presuppositions, a condition that, according to Woollard, would make the fact substantial and, in turn, the action "doing harm".[31] What about Remove and Interpose? Arguably, in Remove the action of the agent is characterised as a substantial fact in the sequence leading to the harmful upshot, and not as mere background conditions: the action thus amounts to a case of doing harm. Interpose, again, may be a case of removing barriers: the barrier here does not belong to the agent nor to the victim, and was not put in place by a third party in order to prevent harm. On Woollard's account, we may argue that the victim has here a stronger non-need claim on the barrier, and thus the action is classified as doing harm.

The main difference between my account and the ones outlined above is that I allow for different reasonable and legitimate expectations and presuppositions, thus accounting for disagreement in these borderline cases. I accept that this account might feel uncomfortable in this respect, as it suggests that, in borderline case, there is no definitive or correct classification

[28] Woollard (2015), p. 23. I also discuss this account in Chapter 1.
[29] Ibidem, p. 79.
[30] Ibidem.
[31] We could also argue that the barrier/respirator does belong to a third party which put the barrier in place in order to prevent harm. Therefore, the removal of the respirator would count again as doing harm.

of actions as doings or else allowings, but different classifications can be tenable depending on our empirical expectations or moral convictions.

Let's start our analysis with Impoverished Village. This case could appear problematic for the self-contained network account, as it looks like the agent, making the phone call, is deviating from what was happening without her intervention, which is thus “more natural” or likely to occur based on the current situation. Nonetheless, following Bennett's intuition, we can argue that the “normal” course of events should be fixed further back in time, before the accountant's unauthorised move. I think that, in this case, the fact that the story begins with what is explicitly defined as a “mistake” explains why most people would fix the normal course of events *before* the mistake was made. Now, we can define the following variables:

D = death of the people in the impoverished village, where Def(D) = 1, since this is what is reasonable to expect if the accountant does not set up the donation;

A = the accountant arranges to make the donation, where Def(A) = 0, since she was neither instructed nor expected to do so;

P = the agent makes the phone call, where Def(P) = 0;

S = the donation is set up, where Def(S) = 0;

These variables can be arranged in a causal graph as follows:

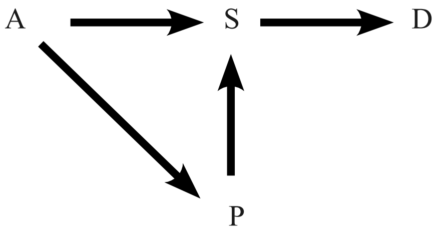


Figure 6.1: Impoverished Village

(the relevant counterfactuals are: D = not-S; S = A & not-P, not-P = not-A).[32]

This interpretation delivers the more reasonable classification in terms of allowing harm, as D can take its deviant value only when A takes its default one, and thus the causal network is not self-contained.

The next three cases, arguably, are more controversial as different expectations about and

[32] These counterfactuals read as: the people in the village would have died if the donation had not been set up; the donation would have been set up if the accountant had arranged for it and the client had not made the phone call; the client wouldn't have made the phone call had the accountant not arranged the donation.

interpretations of what counts as the normal course of events are seemingly more reasonable. In Hospital, the action counts as allowing if we stick again to Bennett's intuition that the normal course of events should be fixed before the doctor's intervention. In this case,

D = death of the patient = 1 if the patient dies, D = 0 if the patient does not die, where Def(D) = 1, as the patient is clearly and immediately going to die unless the respirator is plugged.

We can thus individuate, as parents of D:

P = the respirator is plugged into the patient, where Def(P) = 1, as this amounts to a routine intervention within the duties of a doctor;

K = the respirator is kept plugged in, where Def(K) = 1.

We can then draw the following causal graph:

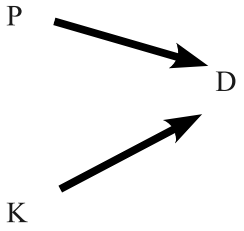


Figure 6.2: Hospital

(the relevant counterfactual is D = P & K).[33]

In this situation, the outcome only takes its default value if P takes its deviant one; the causal network is thus not self-contained and K = 0, the removal of the respirator, counts as allowing harm. Nonetheless, it is not unreasonable for agents to take the scenario *after* the doctor's intervention as the reference point, which will lead to a different causal network with different variables and assignments of default values:

Def(D) = 0, as the respirator is already plugged in and we thus have no reason to expect the patient to die soon.

R = removal of the respirator, the parent of D, would thus take its default value at 0, as this is not standard or expected conduct on the part of the doctor.

In this second causal network, D takes its default value when its parent does; the causal network is thus self-contained and the doctor's action counts as doing harm.

Both Remove and Interpose, arguably, share a similar structure: these actions will be classified as allowings if the normal course of events is set sufficiently far back in the past such that Bill is dying of thirst (Interpose) and Bill is going to be crushed by the cart (Remove). On the other

[33] The patient would have died if the doctor had not plugged the respirator in and kept it plugged in.

hand, if the reference point for what counts as the natural course of events is set when the cart full of water is rolling downhill (Interpose) and the rock is stopping the cart (Remove), these actions would count as doings. Specifically, for the Interpose case, in the allowing interpretation we will set:

D = death by thirst, where $\text{Def}(D) = 1$;

C = a cart full of water rolls towards the victim, where $\text{Def}(C) = 0$, as this is not what we should reasonably expect under normal circumstances;

I = the agent interposes a rock, where $\text{Def}(I) = 0$;

A = a cart full of water reaches the victim,where $\text{Def}(A) = 0$.

These four variables can be arranged in the following graph:

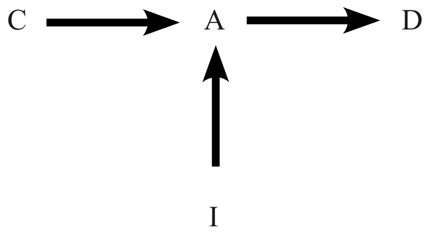


Figure 6.3: Interpose

($A = C$ & not-I, D = not-A, not-I = not-C).[34]

Here, the causal network is not self-contained, as D can take its deviant value when all its parent takes its default one (the cart does not roll towards the victim, and does not reach the victim; the agent does not interpose anything). $I = 1$ is thus characterised as allowing harm.

On the other hand, in the doing interpretation, we have:

$\text{Def}(D) = 0$, as the cart full of water is already rolling downhill;

$\text{Def}(I) = 0$.

The causal network is now self-contained and the action counts as doing harm.

The Remove case can be analysed analogously, leaving room for both interpretations depending on where the reference point for the natural course of events is set.

In conclusion, while for the Smith/Jones example and trolley cases the self-contained network account delivers a “correct” classification in terms of doing rather than allowing, borderline

[34] The cart would have arrived if it had been rolling downhill and the agent had not interposed a rock; the victim would have died if the cart had not arrived.

cases are characterised by an inherent ambiguity, which makes different characterisations legitimate. The ambiguity here appears to be located at the level of selecting from which vantage point we consider the question of what flow of events from *now* on counts as natural or normal. This selection might amount to an empirical matter, and, in borderline cases, both courses of events can reasonably be seen as natural depending on which point in time we fix as the relevant reference point. Plausibly, some reference points will be less natural than others; in Interpose and Remove, for instance, we could question whether, in allowing interpretations, this is set too far back in the past. Again, the more agents converge on a specific interpretation of the course of events, the less we will observe disagreement in doing and allowing attributions.

Conclusion

This thesis has examined the doing/allowing distinction trying to account for two main judgments most of us intuitively endorse when we think about doing *versus* allowing harm. The first is that doing and allowing amount to two distinct ways in which an agent can be causally relevant to a harm. The second is that, all other things being equal, doing harm is harder to justify than allowing harm to occur. Upon closer analysis, however, doing/allowing classifications are frequently contested, and there is persistent disagreement on whether doing is worse than allowing even in cases where making the distinction is uncontroversial. Beyond preserving the two main intuitions, I have thus also suggested that we should strive to explain and account for disagreement and ambiguity.

My proposed account of the doing/allowing distinction is a causal model which relies on a preliminary value assignment to the variables in order to identify which kind of impact an agent has on an outcome. Setting the "default" value for each variable ultimately depends on what we perceive as the "normal course of events". I have argued that "normal" can be interpreted in both an empirical and a normative sense: what we expect will happen and what we expect should happen. So, doing/allowing classifications incorporate descriptive and normative features of cases. In particular, doing/allowing classifications may capture whether an agent violated a standard norm or rule of behaviour, or her conduct was somehow "deviant". This aspect, I suggested, explains why doing/allowing classifications often match up with other morally relevant considerations, and they are therefore morally significant. Nonetheless, which course of events is perceived as natural is ultimately an empirical matter. Agents can have different expectations about what is more likely to happen, about which norms we should conform to, and disagree on how to weigh different principles. As a result, doing/allowing classifications can be context-sensitive and controversial. In "borderline" cases, I claimed that different doing/allowing classifications might be (more or less) reasonable.

This feature of my model could appear unattractive to some. People who argue that our intuitive case judgements on doing/allowing should be taken seriously might indeed worry that this conclusion undermines such intuitions. However, in spite of some ultimately ambiguous cases, our everyday use of the doing/allowing distinction as a guide to moral judgements is still preserved. In most familiar situations, doing/allowing classifications may provide a "composite judgement" which accounts for different features which matters to moral evaluation. Regarding controversial examples, my model may help locating underlying disagreement, and thus hopefully advances discussion.

My proposal also challenges the idea that, because disagreement and framing effects are real, our moral intuitions are never to be trusted. The fact that doing/allowing classifications might be ambiguous does not mean that we should not investigate whether they can nonetheless capture morally relevant features. While consistency in moral theorising is worth pursuing, we may accept that our use of the doing/allowing distinction is complex and nuanced. Therefore, evidence of instability and controversy, especially in under-described and contrived cases, should not rule out the possibility that doing is worse than allowing in familiar cases. My model may thus help in explaining to the sceptic why most cases are, in fact, agreed-upon.

My proposal, finally, still leaves some gaps, and it is not systematic as one might hope. My explanation of disagreement and, more generally, of controversial doing/allowing classifications, relies on a case-by-case analysis. I have suggested that our doing/allowing classifications depend on what agents consider to be the "normal course of events", but I have not offered a detailed account of how this normal course of events should be determined. Specifically, I explained how both empirical and normative expectations may set the default value, but not how the two reconcile in practice. It might be the case that where moral norms are pertinent, these "trump", but if no moral norm is particularly pertinent, we may revert to mere empirical expectations. Further work, however, should be done so as to illuminate on this.

Bibliography

Armstrong, David. M., 1997, *A World of States of Affairs*, Cambridge: Cambridge University Press.

Aronson, Jerrold, 1971, "On the Grammar of 'Cause'", *Synthese*, 22: 414–30.

Barry, Christian, Lindauer, Matthew and Øverland, Gerhard, 2014, "Doing, Allowing, and Enabling Harm: An Empirical Investigation", in J. Knobe, T. Lombrozo and S. Nichols (eds.), *Oxford Studies in Experimental Philosophy*, Oxford: Oxford University Press.

Barry, Christian and Øverland, Gerhard, 2017, *Responding to Global Poverty: Harm, Responsibility and Agency*, New York: Cambridge University Press.

Beebee, Helen, 2004, "Causing and Nothingness", in J. Collins, N. Hall and L. A. Paul (eds.), *Causation and Counterfactuals*, Cambridge, MA: The MIT Press, pp. 291–308.

Bennett, Jonathan, 1980, "Morality and Consequences", *The Tanner Lectures On Human Values*.

Bennett, Jonathan, 1995, *The Act Itself*, Oxford: Clarendon Press.

Bratman, Michael, 1987, *Intentions, Plans and Practical Reasons*, Cambridge, MA: Harvard University Press.

Coady, David, 2004, "Preempting Preemption", in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*, Cambridge, MA: The MIT Press, pp. 325–40.

Cushman, Fiery A., Knobe, Joshua and Sinnott-Armstrong, Walter, 2008, "Moral Judgments Affect Doing/Allowing Judgments", *Cognition,* 108: 281–289.

Donagan, Alan, 1977, *The Theory of Morality*, Chicago: The University of Chicago Press.

Dowe, Phil, 2000, *Physical Causation*, Cambridge: Cambridge University Press.

Dowe, Phil, 2001, "A Counterfactual Theory of Prevention and 'Causation' by Omission", *Australasian Journal of Philosophy*, 79: 216–26.

Ducasse, C. J., 1926, "On the Nature and Observability of the Causal Relation", *Journal of Philosophy*, 23: 57–68.

Dworkin, Ronald, Nagel, Thomas, Nozick, Robert, Rawls, John, Scanlon, Thomas and Thomson, Judith Jarvis, 1997, “The Philosopher's Brief”, *The New York Reviews of Books*, 1997 (27): 41–47.

Elster, Jakob, 2011, “How Outlandish Moral Cases can be?”, *Journal of Applied Philosophy*, 28 (3).

Fitzpatrick, William J., 2006, “The Intend/Foresee Distinction and the Problem of ‘Closeness’”, *Philosophical Studies,* 128: 585–617.

Foot, Philippa, 1978, “The Problem of Abortion and the Doctrine of Double Effect”, in *Virtues and Vices and Other Essays*, Berkeley, CA: University of California Press.

Foot, Philippa, 1984, “Killing and Letting Die”, in J. L. Garfield and P. Hennessey (eds.), *Abortion: Moral and Legal Perspectives*, Amherst: University of Amherst Press, pp. 355–382.

Foot, Philippa, 1985, “Morality, Action and Outcome”, in T. Honderich (ed.), *Morality and Objectivity*, London: Routledge and Kegan Paul.

Greene, Joshua D., 2015, “Solving the Trolley Problem”, in J. Systma and W. Buckwalter (eds.), *A Companion to Experimental Philosophy*, John Wiley & Sons Ltd, Chapter 11.

Greene, Joshua D., 2015, “The rise of Moral Cognition”, *Cognition*, 135: 39–42.

Greene, Joshua D., Cushman, Fiery A., Stewart, L. E., Lowenberg, K., Nystrom, L. E. and Cohen J. D., 2009. “Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment”, *Cognition*, 111(3): 364–371.

Halpern, Joseph Y. and Hitchcock, Christopher, 2013, “Graded Causation and Defaults”, *The British Journal for the Philosophy of Science*, 0 (2013): 1–45.

Halpern, Joseph. Y. and Hitchcock, Christopher, 2010, “Actual causation and the art of modeling”, *Causality, Probability, and Heuristics: A Tribute to Judea Pearl*, London: College Publications, pp. 383–406.

Halpern, Joseph Y. and Pearl, Judea, 2000, “Causes and Explanation: a Structural-Model Approach. Part II: Explanations”, *The British Journal for Philosophy of Science*, 56(2005): 899–911.

Hitchcock, Christopher, 2001, “The Intransitivity of Causation Revealed in Equations and

Graphs", *Journal of Philosophy*, 98: 273–99.

Hitchcock, Christopher, 2007, "Prevention, Preemption, and the Principle of Sufficient Reason", *Philosophical Review*, 116: 495–532.

Hitchcock, Christopher, 2009, "Cause and Norm", *The Journal of Philosophy*, Volume cvi, 11: 587–612.

Horgan, Terry and Timmons, Mark, 2009, "What does the Frame Problem Tell Us About Moral Normativity?", *Ethical Theory and Moral Practice*, 12(1): 25–51.

Horowitz, Tamara, 1998, "Philosophical Intuitions and Psychological Theory", *Ethics,*108: 367–85.

Howard-Snyder, Frances, 2002, "Doing vs. Allowing Harm," *The Stanford Encyclopedia of Philosophy* (Summer 2002 Edition) Edward N. Zalta (ed) URL = <https://plato.stanford.edu/archives/sum2002/entries/doing-allowing>

Jou, Jerven, Shanteau, James and Harris, Richard J., 1996, "An information processing view of framing effects: The role of causal schemas in decision making", *Memory & Cognition*, 24 (1): 1–15.

Kahneman, Daniel, Knetsch, Jack L. and Thaler, Richard H., 1990, "Experimental Tests of the Endowent Effect and the Coase Theorem", *Journal of Political Economy*, 98(6): 1325–1348.

Kahneman, Daniel and Tversky, Amos, 1979, "An Analysis of Decision under Risk", *Econometrica*, 47(2): 263–291.

Kahneman, Daniel and Tversky, Amos, 1983, "Choice, Values and Frames", *American Psychologist*, 39(4): 341–350.

Kagan, Shelly, 1989, *The Limits of Morality*, Oxford: Oxford University Press.

Kamm, Frances, 1996, *Morality, Mortality*, Volume II, Oxford: Oxford University Press.

Kamm, Frances, 1998, "Moral Intuitions, Cognitive Psychology and the Harming-versus-Not-Aiding Distinction, *Ethics*, 108(3): 463–288

Kamm, Frances, 2007, *Intricate Ethics*, New York: Oxford University Press.

Kistler, Max, 1998, "Reducing Causality to Transmission", *Erkenntnis*, 48: 1–24.

Knobe, Joshua, 2006, “The concept of intentional action: a case study in the uses of folk psychology”, *Philosophical Studies*, 130: 203–231.

Levin, Irwin P., Schneider, Sandra L., and Gaeth, Gary J., 1998, “All frames are not created equal: a typology and critical analysis of framing effects”, *Organizational Behavior and Human Decision Processes*, 76: 149–188.

Lewis, David, 1986, “Causation,” in his *Philosophical Papers* (Volume 2), New York: Oxford University Press.

List, Christian and Gold, Nathalie, 2004, “Framing as Path-Dependence”, *Economics and Philosophy*, 20(02): 253–277.

Mackie, John, 1974, *The Cement of the Universe*, Oxford: Oxford University Press.

Mangan, Joseph T., 1949, “An Historical Analysis of the Principle of Double Effect”, *Theological Studies*, 10: 41–61.

McMahan, Jeff, 1993, “Killing, Letting Die and Withdrawing Aid”, *Ethics*, 103 (January): 250–279.

McMahan, Jeff, 1998, “Review: A Challenge to Common Sense Morality”, *Ethics*, 108(2): 394–418.

McMahan, Jeff, 2000, in Hugh LaFollette (ed.), *Blackwell Guide to Ethical* Theory, Oxford: Blackwell.

McMahan, Jeff, 2018, “Torture and Methods in Moral Philosophy”, in S. Anderson and M. Nussbaum (eds.), *Confronting Torture: Essays on the Ethics, Legality, History, and Psychology of Torture Today*, Chicago: University of Chicago Press.

Menzies, Peter, 2003, “Difference-Making in Context”, in J. Collins, N. Hall and L. Paul (eds.), *Counterfactuals and Causation*, Cambridge, MA: MIT Press.

Menzies, Peter, 2004, “Causal models, Token Causation, and Processes”, in *Philosophy of Science*, 71(5): 820–32.

Menzies, Peter and Price, Huw, 1993, “Causation as a Secondary Quality”, *British Journal for the Philosophy of Science*, 44(2): 187–203.

Nelkin, Dana K. and Rickless, Samulel C., 2005, “So Close,Yet So Far. Why Solutions to the

Closeness Problem for the Doctrine of Double Effect Fall Short", *Nous,* 49:2 (2015): 376–409.

Osman, Magda, 2014, "Dynamic Moral Judgement", *Psychology*, 2014, 6, Published Online June 2015 in SciRes. http://www.scirp.org/journal/psych

Paul, L. A., 2000, "Aspect Causation", *Journal of Philosophy*, 97: 223–34.

Pearl, Judea, 2000, *Causality*, Cambridge: Cambridge University Press.

Persson, Johannes, 2002, "Cause, Effect, and Fake Causation", *Synthese*, 131: 129–143.

Petrinovich, Lewis and O'Neill, Patricia, 1996, "Influence of Wording and Framing Effects on Moral Intuitions", *Ethology and Sociobiology*, 17: 145–171.

Petrinovich, Lewis, O'Neill, Patricia, and Jorgensen, Matthew, 1993, "An empirical study of moral intuitions: Toward an evolutionary ethics"; *Journal of Personality and Social Psychology,* 64(3), 467–478.

Pettit, Philip, 2015, *The Robust Demand of the Good: Ethics with Attachment, Virtue, and Respect*, Oxford: Oxford University Press.

Quinn, Warren S., 1989, "Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing", *Philosophical Review*, 98 (3): 287–312, reprinted in Steinbock and Norcross 1994, pp. 355–382.

Rachels, James, 1975, "Active and Passive Euthanasia", *New England Journal of Medicine*, 292: 78–86.

Rickless, Samuel C., 2011, "The Moral Status of Enabling Harm", *Pacific Philosophical Quarterly*, 92(1): 66–86.

Rickless, Samuel C., 1997, "The Doctrine of Doing and Allowing", The Philosophical Review, 106(4): 555–575.

Salmon, Wesley, 1984, *Scientific Explanation and the Causal Structure of the World*, Princeton: Princeton University Press.

Sartorio, Carolina, 2005, "Causes as Difference-Makers", *Philosophical Studies*, 123: 71–96.

Schaffer, Jonathan, 2000, "Causation by Disconnection", *Philosophy of Science*, 67: 285–300.

Schaffer, Jonathan, 2005, “Contrastive Causation”, *The Philosophical Review*, 114(3): 297–328.

Schaffer, Jonathan, 2016, “Grounding in the image of causation”, *Philosophical Studies*, 173: 49–100

Shafir, Eldar, and LeBoeuf, Robyn A., 2002, “Rationality”, *Annual Review of Psychology*, 53: 491–517.

Scheffler, Samuel, 2005, “Doing and Allowing”, *Ethics*, 114(2): 215–239.

Sher, Shlomi and McKenzie, Craig Rm, 2008, “Framing Effects and Rationality”, in Chater and Oaksford (eds.), *The Probabilistic Mind: Prospects for Rational Models of Cognition*, Oxford: Oxford University Press.

Singer, Peter, 1979, *Practical Ethics*, Cambridge: Cambridge University Press.

Sinnott-Armstrong, Walter, 2005, “Framing Moral Intuitions”, in W. Sinnott-Armstrong (ed.), *Moral psychology, Vol. 2. The cognitive science of morality: Intuition and diversity,* pp. 47-76, Cambridge, MA: MIT Press.

Sinnott-Armstrong, Walter, Mallon, Ron, McCoy, Tom and Hull, Jay G., 2008, “Intentions, Temporal Order and Moral Judgments”, *Mind and Language*, 23 (1): 90–106.

Skyrms, Brian, 1984, “EPR: Lessons for Metaphysics,” in P. French, T. Uehling, Jr., and H. Wettstein (eds.), *Midwest Studies in Philosophy IX*, Minneapolis: University of Minnesota Press, pp. 245–55.

Sunstein, Cass R., 2003, “Moral Heuristics”, *Univeristy of Chicago Law & Economics*, Oline Working Paper No. 180.

Tadros, Victor, 2015, “Wrongful intentions without closeness”, *Philosophy & Public Affairs*, 43(1): 52–74.

Thomson, Judith Jarvis, 1986, “Killing, Letting Die and the Trolley Problem”, *Rights, Restitution, and Risk: Essays in Moral Theory*, W. Parent (ed.), Cambridge: Harvard University Press.

Thomson, Judith Jarvis, 2008, “Turning the Trolley”, *Philosophy and Public Affairs*, 36 (4): 359–374.

Tooley, Michael, 1972, “Abortion and Infanticide”, *Philosophy and Public Affairs*, 2 (1): 37–65.

Trammell, Richard, 1975, “Saving and Taking Life”, *The Journal of Philosophy*, 72: 131–137.

Trammell, Richard, 1979, “The Nonequivalency of Saving Life and Not Taking Life”, *The Journal of Medicine and Philosophy*, 4 (3): 251–262.

Tversky, Amos, and Thaler, Richard H. (1990). Preference reversals. Journal of Eco- nomic Perspectives 4: 201–211.

Von Wright, G. H., 1975, “On the Logic and Epistemology of the Causal Relation,” in *Causation and Conditionals*, Oxford: Oxford University Press, pp. 95–113.

Wolff, Philip, 2007, “Representing Causation”, *Journal of Experimental Psychology*: 136(1): 82–111.

Woodward, James, 2003, “Making Things Happen: A Theory of Causal Explanation”, Oxford Oxford University Press.

Woodward, James, 2004, “Counterfactuals and Causal Explanation”, *International Studies in the Philosophy of Science,*18: 41–72.

Woodward, James and Hitchcock, Christopher, 2003, “Explanatory Generalizations, Part I: A Counterfactual Account”, *Noûs*, 37(1): 1–24.

Woollard, Fiona, 2008, “Doing and Allowing, Threats and Sequences”, *Pacific Philosophical Quarterly*, 89: 261–277.

Woollard, Fiona, 2013, “If this is My Body...: a Defence of the Doctrine of Doing and Allowing”, *Pacific Philosophical Quarterly*, 94: 315–341.

Woollard, Fiona, 2015, *Doing and Allowing Harm*, Oxford: Oxford University Press.