London School of Economics and Political Science

Agency as Difference-making

Causal Foundations of Moral Responsibility

Johannes Himmelreich

A thesis submitted to the Department of Philosophy, Logic and Scientific Method of the London School of Economics for the degree of Doctor of Philosophy, London, July 2015

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 (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of 59,009 words.

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 Ryan Cox, Natalie Cutting, Jesse Saloom, and Ying Shi.

Abstract

We are responsible for some things but not for others. In this thesis, I investigate what it takes for an entity to be responsible for something. This question has two components: agents and actions. I argue for a permissive view about agents. Entities such as groups or artificially intelligent systems may be agents in the sense required for responsibility. With respect to actions, I argue for a causal view. The relation in virtue of which agents are responsible for actions is a causal one. I claim that responsibility requires causation and I develop a causal account of agency. This account is particularly apt for addressing the relationship between agency and moral responsibility and sheds light on the causal foundations of moral responsibility.

Acknowledgements

There is a difference between doing something yourself and something being up to you. Responsibility requires the latter but not the former. This is a central claim of this thesis and it applies to itself. I have done the writing myself, but what I did was not all up to me. Numerous individuals and institutions have helped me complete this thesis. I thank each of them and apologise to anyone I have forgotten to mention.

My supervisors, Christian List and Richard Bradley, have been invaluable to me with their feedback, encouragement, and advice. I have learned a great deal from each of them and have thoroughly enjoyed the excitement and intellectual spirit at the London School of Economics' (LSE's) Philosophy department and the Choice Group in particular.

I have benefitted from valuable comments and discussions at the following events: the Graduate Conference at the Universidade Nova de Lisboa in 2012; the 8th Meeting of the German Analytic Philosophy Association (GAP.8) in Konstanz in 2013; the meetings of the Australasian Association of Philosophy in Brisbane and Canberra in 2013 and 2014 respectively, as well as the Philsoc and the MSPT seminars at the Australian National University (ANU) around the same time; the 3rd Conference of the European Network on Social Ontology and the Philosophy of Science seminar at the TINT Centre of Excellence at the University of Helsinki in 2013; the University of London graduate conference; the meetings of the Moral Attitudes and Decision-making Lab (Madlab) at Duke University; the LSE Workshop on The Politics and Philosophy of Artificial Intelligence (AI); the Manchester Workshop on Collective Obligations; the Collective Intentionality IX meeting in Bloomington, Indiana; the Princeton Political Theory seminar, and the American Philosophical Association (APA) Eastern Division Meeting in Philadelphia, all of the latter which took place in 2014. In 2015 some of the work in this thesis was presented at the "Doktorandenforum" workshop of the Studienstiftung des Deutschen

Volkes in Heidelberg; at a workshop on moral motivation in Sheffield; and at the VI. Meeting on Ethics and Political Philosophy in Braga, Portugal. I am furthermore grateful for the opportunities to present my work at and feedback from the LSE Choice Group in 2012 and at the various sessions of LSE's philosophy PhD student seminar.

I am very grateful for the financial support I have received from LSE, and the academic environment and inspiration provided by the German Academic Merit Foundation (Studienstiftung des Deutschen Volkes), the Princeton University Center for Human Values, and the Madlab at Duke University. The ANU's School of Philosophy have been especially welcoming and hospitable during each of my visits. I have benefitted hugely from the Coombs' environment and its wonderful people.

My personal thanks go to Walter Sinnott-Armstrong, Holly Lawford-Smith, Daniel Stoljar, Nic Southwood, Colin Klein, Daniel Nolan, Seth Lazar, LA Paul, Christopher Thompson, Foad Dizadji-Bahmani, Ben Ferguson, Orri Stefánsson, Jason McKenzie Alexander, Gabriel Wollner, Matthew Braham, Alma Barner, Philip Pettit, Andreas Tupac Schmidt, and Sebastian Köhler. Marco Mayer has been an invaluable inspiration and interlocutor for many years. He scrutinised central ideas of this thesis with me before I even began putting them in writing. His comments have helped me avoid many mistakes. Jesse Saloom has been a constant and loyal support with his comments, encouragement, and friendship. His rigour and sense of style are exceptional and I am glad for every hour he has spent on my thesis. Ryan Cox has commented on almost all parts of this thesis at various stages. He has drawn my attention to many issues I would have overlooked. I have learned much from him. Ying Shi's contribution cannot be overstated. My writing has benefitted enormously from her acute sense for removing redundancies, "technical terms", and other jargon. Without her, not only would this thesis be harder to read, but I may not have finished some parts of it at all.

Every line in this thesis has benefitted from the love and support of my parents. My gratitude to them is beyond words.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

What does it take for someone to be responsible for something? This is the topic of this thesis. I focus on whether moral responsibility requires causation. I argue that agents are responsible only for the things they cause, and in the process, I develop causal foundations for moral responsibility. My proposal is novel, controversial, and simple.

In this introduction I motivate and clarify the topic. The thesis itself consists of several chapters, some of which are explorative, while others are constructive. The former probe initial restrictions and assess sceptical arguments. The latter, in the second part of the thesis, put forward the proposal, spell out its details, and discuss examples.

This introduction comes in three parts. In Section 1.1 I explain how I understand the subject of the thesis. In Section 1.2 I discuss two cases to provide an overview of the topic. I point out in what way the answer I give is novel. In Section 1.3 I state and justify assumptions that I take for granted and I explain why the proposal is controversial. Towards the end of the thesis, in Section 8.2, I reflect on my method. There, I identify the sense in which my proposal is simple and why that matters.

1.1 The Metaphysics of Moral Responsibility

What does it take for someone to be responsible for something? In part, this is a question of moral metaphysics. It concerns the atoms and elements constituting the moral matter of our world. What are these atoms? And what are the bonds and relations between them? In my view, the atoms of moral philosophy are agents and their actions. The crucial bond between them is the causal relation. Agents are responsible for actions partially in virtue of causing them. There is no responsibility without causation. This claim has recently been met with objections pointing to problems with theories of causation (cf. Sartorio, 2004, 2007). I put forward a novel proposal concerning the causal bond between agents and actions that solves some of these problems.

Responsibility is at the heart of moral philosophy. Many legal doctrines and social practices rest on it. Moral responsibility in turn rests on the concept of agency and the metaphysics of causation. The subject of this thesis is located in this range between the metaphysics of causation and legal doctrines. It involves a distinct set of research topics, each of which is related in its own way to moral responsibility.

When we hold someone responsible, we take three aspects into account. First, we judge a person for what she has done. Responsibility has an evaluative aspect. Second, we draw conclusions when we find someone responsible. Someone's responsibility gives us a reason to shape our interactions with them. Responsibility has a reason-giving aspect. Third, we resent or we esteem those who we hold responsible. Responsibility has an emotional aspect. In this thesis I investigate what it takes for someone to be responsible. Specifically, I focus on one particular kind of necessary condition for moral responsibility: agency.

Agency is a necessary condition for responsibility. I am responsible for some actions but not for others. Generally, I am responsible for what I do; you are responsible for what you do. This is because you are an agent of your actions, and I am an agent of mine. This thesis is about spelling out this thought and vindicating it. It may sound simple, but in some situations it is not obvious who is responsible for what. To illustrate this point, consider two cases.

1.2 Two Cases

In the following two cases you will find most of the issues that, taken together, form the topic of this thesis. Each case raises the question: *who is responsible?* I present each case, briefly answer this question, and then raise concerns regarding the answer. My aim is to explain my motivation for the topic and to clarify the connections between chapters by pointing to the parts of the thesis where I discuss these concerns.

1.2.1 A Vertical Group

> **Commanded Killing**. Suppose Anne, a military commander, commands her team to shoot an innocent civilian, Collin. The team consists of two soldiers, Bert and Ben. Bert usually shoots first. But if he were not to shoot, then Ben would. Bert goes ahead and shoots Collin.

This case is a model for hierarchical or *vertical groups*. In some features it resembles private corporations, government agencies, and of course the military. We also find hierarchical groups at work in history. For example, up until the autumn of 1989, when Germany was divided into East and West, East German soldiers at the Berlin Wall had orders to shoot fugitives. In Berlin alone, 67 people were shot while trying to escape (Hertle and Nooke, 2009, 18–25).

Who is responsible? Following Germany's reunification, the border soldiers were prosecuted and convicted for manslaughter and murder. But the courts did not only convict soldiers; they also convicted military and party leaders. This coincides with our intuitive judgments. Responsibility falls on those who give orders as well as on those who follow them. Or so we believe.

The literature currently lacks a philosophical account of responsibility in hierarchical groups. My thesis aims to rectify this. I examine answers to the following question: How do we support the claim that Anne is responsible as well as Bert, or that East German commanders are responsible as well as their soldiers?

The challenge is to understand what kind of agency is involved in responsibility. Let us consider two principles about the kind of agency that is necessary

for responsibility. The first principle says that responsibility requires performance. An individual is responsible for an action only if she *performs* the action in question. When applied to our case, it says that the soldier *is* responsible, but it implies that the commander is *not* responsible. The soldier performs the killing but the commander does not. This seems inconsistent with our intuitions. Intuitively, we believe the commander *is* responsible for the killing. Therefore, this principle seems false.

The second principle says that responsibility requires dependence. An individual is responsible for an action only if the action *would not have occurred* were it not for the individual. This time, the picture is reversed. Applied to our case, this condition says that the commander *is* responsible, but it implies that the soldier is *not* responsible. Collin's death depends on the commander but not on Bert. Had it not been for the commander, Collin would not have died. But no matter what Bert had done, Collin would still have died. If Bert had not fired, then Ben would have. Again, this seems inconsistent with our intuitions. Intuitively, we believe the soldier *is* responsible for the killing. Therefore, this principle seems false.

We hold the soldier responsible but not the commander by the first principle, while we hold the commander responsible but not the soldier by the second. Neither vindicates the thought that both agents — soldier and commander — are responsible. In the case of both principles, at most one agent is responsible. Should we accept both principles? Or should we give up one of our beliefs about who is responsible?[1] The aim of this thesis is to find a single principle of responsibility, while respecting important differences between the soldier and the commander. I argue that responsibility requires difference-making.

I put forward an account of agency in hierarchical groups in Chapter 4. At the centre of this account is the notion of *agency as difference-making*, which I spell out in Chapter 5. More generally, I describe and defend this proposal in detail during most of the chapters of Part II. It relies on the distinction between being an agent of an action and performing an action. I develop this distinction in Chapter 3.

[1] A further option is to say that there are two kinds of responsibility rather than one, each associated with one of these principles. The commander would be responsible in one sense, and the soldier would be responsible in a different sense. Fischer and Ravizza (2000) seem to propose such a dichotomous solution. They distinguish between regulative-control and guidance-control. See also Sartorio (2007).

1.2.2 A Horizontal Group

Let us now look at a second case that concerns an alternative kind of organisation.

> **Collective Decision**. An investment committee decides whether or not to fund a start-up. All committee members agree that the start-up should be funded if and only if three conditions hold: first, the business idea is promising; second, the budget plan is sound; and third, the founder's track record is strong. As it happens, each member individually finds fault with the budget plan, the business idea, or the founder's track record. Therefore no individual believes that the start-up should be funded. The committee votes on the three propositions. The business idea, the budget plan, and the track record each gets a majority of the members' votes. The head of the committee signs the cheque on this basis. As it turns out, the start-up closes down a year later and the investment is lost.

The situation seems paradoxical (List and Pettit, 2002). On the one hand, no individual believes that the start-up should be funded. On the other hand, there is a majority supporting all the reasons for doing so. Such situations arise in practice. The case of Collective Decision is a model for *horizontal groups*. Members in horizontal groups share a common goal: they cooperate and react to each other, and influence an outcome on a roughly equal footing.[2]

Who is responsible? There are three responses to this question: everybody, nobody, or the head of the committee. Let us look at the first response. According to the first response, every individual on the committee is responsible for the bad investment. This is because each member contributes to the collective decision. Moreover, the outcome of the decision might have

[2]Each of these notions can be spelled out. Bratman (2014) discusses some of these features under the heading of "modest sociality". The relevant sense of "influence" here is that of *ex ante* influence. A member's influence can be measured *ex ante*, that is, independent of the choices of other members, and *ex post*, that is, given the choices of other members (Lagnado et al., 2013).

been different if any individual had voted differently.[3] Let us call this the *individual responsibility* response.

But there are problems with the individual responsibility response. First, in most cases with more than three members, the outcome of the decision would not change if any individual had voted differently.[4] Second, the group decides as a collective entity. The members together believe that the start-up should be funded. As a group they believe that the business idea is promising, that the budget plan is sound, and that the track record is strong. In fact, on his or her own, each member of the investment committee intends *not* to fund the start-up. And yet, the start-up is funded. This is not an accident; it is an intentional action. Assuming that there must be someone who intends to fund the start-up, it makes sense to say that it is the members of the committee who fund the start-up, not individually, but as a group.

The second response agrees with this argument. It says that responsibility is situated not at the individual level, but rather at the group level. The group is an entity over and above the committee members and it is responsible for the collective action of funding the start-up. Let us call this the *collective responsibility* response.

Many find the collective responsibility response hard to accept. Our intuitions and emotional reactions favour pointing a finger at someone. We seem to believe that for any action there must be an individual responsible for it.[5] This rules out collective responsibility. But there is also intuitive support in favour of collective responsibility. Consider the following passage from John Steinbeck's (1939) *The Grapes of Wrath*, on foreclosure.

> Sure, cried the tenant men, but it's our land. We measured it and broke it up. We were born on it, and we got killed on it, died on it. Even if it's no good, it's still ours. That's what makes it ours — being born on it, working on it, dying on it. That makes ownership, not a paper with numbers on it.

[3] In contrast to the situation in the Commanded Killing case, both the performance and the dependence principles mentioned earlier are met. On the performance principle, each individual performs his or her part. It seems true in this case that an individual is responsible insofar as she performs the action in question. The second principle also seems to be satisfied. This is the principle that states that an individual is responsible only if, had it not been for this individual, the action would not have occurred.

[4] For example, if the committee had six members and each proposition had the support of four members, then no individual member could change the outcome of the vote.

[5] This principle is not unreasonable. The argument that driverless cars will lead to a so-called responsibility gap might implicitly appeal to this principle.

> We're sorry. It's not us. It's the monster. The bank isn't like a man.
>
> Yes, but the bank is only made of men.
>
> No, you're wrong there — quite wrong there. The bank is something else than men. It happens that every man in a bank hates what the bank does, and yet the bank does it. The bank is something more than men, I tell you. It's the monster. Men made it, but they can't control it.

In addition to holding *each* individual responsible and holding *no* individual responsible, there is a third response. This third response says that the head of the committee is responsible. After all, she is the one who performs the action that directly causes the bad investment.[6] She signs the cheque and transfers the money that is then lost. Let us call this the *enactor responsibility* response.

There is an obvious problem with the enactor responsibility response. The head only acts on behalf of the group. When she signs the cheque and transfers the money she does so only because the bank requires her to. Similarly to how a lawyer represents a client, and a bailiff carries out a court order, the head represents the group and carries out the group's decision in signing the cheque. Think again of the bank employees in the Steinbeck passage. They have a point in saying that they merely carry out what some other entity — "the monster" — requires them to do.

I discuss each of these three responses at length in the thesis. To foreshadow my conclusion, I favour the collective responsibility response. I argue that the decision to fund the start-up is a collective action for which the group as a whole is responsible. Responsibility requires that an action is appropriately related to the mental state of the agent. I contend that the group as a whole stands in this relation to the collective action of funding the start-up.

The argument in support of collective responsibility extends over almost all chapters in this thesis. The positive part of the argument is in Chapter 5. In this chapter I develop in detail the relation that responsibility requires. I argue that the relation in virtue of which agents are responsible for an action is the *difference-making* relation. An agent is related appropriately to an action

[6] I want to register my discomfort with the phrase "directly causes". I suspect that to analyse this notion, we would need to appeal to a particular sort of theory about causation. However, this phrase is used in the literature. See, for example, Ludwig (2007a).

when she has an intention that makes a difference to the occurrence of the action.

There is also a negative part of the argument in support of collective responsibility. It rejects the alternative answers of individual responsibility and enactor responsibility. In Chapter 2 I develop the objection to individual responsibility. It states that the individuals cannot be responsible because the committee acts as a group. I argue that collective actions exist in the sense that committees can act as a group, and that a widespread argument for reducing collective actions to individual actions fails. In Chapter 3 I develop the objection to enactor responsibility. It states that the head of the committee cannot be responsible because she acts only on behalf of the group. I formulate the conditions for performing an action on behalf of another agent and contrast this with other forms of agency in which an agent herself remains inactive as far as overt bodily movements are concerned.

1.3 Basic Assumptions

During the entire thesis I take two things for granted. First, I assume that agency is a relation. Second, I assume that responsibility requires agency. With the first assumption, I set the focus on one particular aspect of the many meanings of "agency". With the second assumption, I formulate a requirement to serve as a test. I want to investigate how far our theories of responsibility can get with this assumption.

1.3.1 Agency as a Relation

My first assumption posits agency as a relation. There are two components: an agent and an action. The agency relation describes which particular entity is an agent of which particular action. This is how I understand "agency" in this thesis.

A different sense of "agency" understands agency as a predicate of an entity. It describes which particular entity is an agent. In contrast to a relation, any element of a predicate includes only one thing. In this thesis, I do not discuss agency as a property that some entities have and that others lack.

This is only meant as a simplification to focus on one particular aspect of the many meanings of "agency" without loss of generality. There is one way,

however, in which this assumption may be seen as a substantive restriction. Agency in the predicative sense may require something other than agency in the relational sense of being an agent of an action. For example, being an agent might require rationality. Indeed, some argue that being rational is constitutive of agency. My proposal is neutral on the question as to what agency is in the predicative sense. We could explicitly disambiguate the use of "agency" between its relational and its predicative sense. Nothing but convenience would be lost.

1.3.2 Responsibility Requires Agency

My second basic assumption is a desideratum. It is a principle that any theory of responsibility should satisfy. I take it for granted that moral responsibility requires agency.

> *Responsibility Requires Agency*.
> For any a and x, if a is responsible for x, then a is an agent of x.

Whether or not this principle is satisfied depends crucially on how we define "agency". The aim of this thesis is to defend a specific definition of "agency" by showing that it satisfies this principle. Of course, we must not restrict our understanding of "agency" only to those actions that an agent performs herself. If we did this, the principle would be implausible. Rather, I want to explore whether there is a way of understanding agency that is broad enough to satisfy this principle. In other words, I do not build *on* this principle but rather I build something *with* it. I take the principle for granted and aim to develop a theory that is independently attractive and that satisfies this principle. I believe that this cause is worth its effort. That responsibility requires agency is a central principle in moral philosophy. I discuss what is at stake if we consider giving it up in Sections 4.1 and 5.1. In particular, my view is that this assumption helps to make our theory of responsibility simple.

In one way, this assumption is uncontroversial. Almost everybody agrees that responsibility requires agency in some sense. Bernard Williams (1994, 55) argues that Homer's works already indicate this much. From investigating the *Iliad* and the *Odyssey* he concludes that there are four "basic elements of any conception of responsibility", one of which is that someone is responsible "in virtue of what he did". John Mackie (1977, 208) calls a similar principle

"the straight rule of responsibility: an agent is responsible for all and only his intentional actions". Likewise, Gideon Rosen (2015, 75) writes that "to resent a person X for an act A is to think, inter alia, that X did A. When you resent Bob for stepping on your foot, you take it *that Bob stepped on your foot*. That's why your resentment evaporates when you discover otherwise (if you are rational)".

In another way, however, this assumption is controversial. Suppose we assume in addition that agency requires causation. Most theories in contemporary philosophy of action assume that agency requires causation.[7] Effectively, we then assume that responsibility requires causation. If someone is responsible for something, then she is a cause of it. But many reject the notion that responsibility requires causation. They must deny either that responsibility requires agency or that agency requires causation. But both — the view that agency requires causation as well as the "straight rule of responsibility" — are central tenets of contemporary philosophy. What then is the relation between responsibility and causation?

Some people deny that responsibility requires agency. Instead, one alternative view suggests that responsibility is only *grounded* in causation but it does not require it (Sartorio, 2004, 2007). I disagree. I argue that responsibility requires causation. My aim is to investigate, against recent arguments, how far we get in assuming that responsibility requires agency. I suggest that we can get quite far. This is the sense in which my proposal is controversial.

Yet, my proposal and this alternative view disagree only in part. Two relations describe the connection between responsibility and causation. On the one hand there is the *entailment* relation. On the other hand there is the *grounding* relation.[8] My proposal and the alternative view disagree about entailment, but agree about grounding.

My proposal not only salvages two important tenets of contemporary philosophy, but with its simplicity it also presents a methodological improvement.[9]

[7] They are causal theories of action. They differ about what exactly the relevant causes involved in agency are. Some argue that the agent herself is a cause. Others argue that the agent's mental states are a cause. I discuss this alleged contrast in Chapter 7.

[8] Entailment is a relation between propositions. It gives us a formal way of understanding the necessary condition expressed by "only if". Grounding is a relation between facts. It describes the direction of metaphysical explanation, that is, it describes which facts are more fundamental than others.

[9] We have well-established formal systems to define entailment and causation. This is not the case for grounding. Even if there were a well-established formal system to define grounding, it would be less simple than the systems needed to define entailment. Unlike entailment, grounding is what Nolan (2014a) calls a hyperintensional notion. That is, ex-

Simplicity is a reason to favour one theory over another. This is the methodological attitude that motivates many arguments in this thesis. I say more about this attitude in Section 8.2 towards the end of the thesis.

1.4 Chapter Abstracts

Chapter 2 The Paraphrase Argument

This chapter is about the status of collective actions. According to one view, sentences about collective actions are merely a shorthand for sentences about individual actions. This is taken to support the claim that collective actions metaphysically reduce to individual actions. I reconstruct this argument and show via counterexamples that it is not sound. The argument relies on a paraphrase procedure to unpack alleged shorthand sentences about collective actions into sentences about individual actions. I argue that the best paraphrase procedure that has been put forward so far fails to produce adequate results.

Chapter 3 Agency and Embodiment

Most philosophers assume that agency must be embodied. In this chapter I argue to the contrary that an agent need not be embodied in order to act. This conclusion has important implications for questions of responsibility and collective agency. It furthermore clarifies how agency works in groups, human–machine interactions, and virtual realities. I begin by analysing the notion of embodiment and presenting three assumptions that comprise this view. First, an agent needs to perform an action herself. Second, the performance of an action requires a bodily movement. Third, a bodily movement involves the movement of a biological body. I deny each of these assumptions. With each assumption that is given up, an additional kind of disembodied action becomes available. The different kinds of disembodied action include proxy actions, extended actions, and extended movements. We are encountering some of these possibilities now, and will plausibly encounter the rest in the near future.

pressions necessarily co-extensive to those involving grounding are not freely substitutable without changing their truth value.

Chapter 4 Responsibility in Hierarchical Groups

In this chapter I lay out a formal account of responsibility in hierarchical groups using the particularly pressing example of future autonomous military drones. Drawing on recent advances in the study of causation, I defend command responsibility. That is, individuals higher up in the chain of command are responsible for actions that subordinates perform on their behalf. I develop a control-based account of agency to explain, on the one hand, how a commander can be held responsible for the bombing that the drone performs, and on the other hand, how a drone exercises agency and counts as autonomous despite the commander's control.

Chapter 5 Agency as Difference-making

This chapter puts forward an account of agency as difference-making. The proposal is that an a is an agent of an action x if and only if a's having a certain intention makes the difference as to whether or not x occurs. I formulate this proposal using a propositional semantics and argue that it is plausible. I then illustrate the usefulness of this account using different cases.

Chapter 6 Omissions

In this chapter I put forward a new solution to the *problem of profligate omissions*. The problem is that some definitions of causation identify any omission that could have prevented an effect as a cause, which leads to counterintuitive results. The solution is to strengthen the counterfactual dependence condition of causation and to weaken the centering condition of the semantics. In contrast to existing solutions, this new solution does not appeal to normative, epistemic, pragmatic, or metaphysical considerations.

Chapter 7 Activity and Exclusion

This chapter offers a reconciliation between two seemingly opposing accounts of agency, namely, the so-called agent-causal account and the event-causal account. The latter faces the *problem of agential activity*, or so argue proponents of the former. Event-causal accounts fail to account for the fact that an agent is active when performing an action. I argue that this problem is an exclusion problem, similar to a well-known argument in philosophy of mind. The two

accounts of agency describe causation on different levels and there need not be any rivalry between them. I clarify the analogy between the two problems and suggest alternative solutions.

Part I

Challenging Individual and Embodied Agency

Chapter 2

The Paraphrase Argument

Are there collective actions? Or are there only actions of individuals, of which we sometimes speak as if they were collective actions? A widely held view, which I call the *paraphrase argument*, answers the latter question in the affirmative. Proponents of this view contend that all statements about collective actions are *merely a shorthand* for statements about the actions of individuals. This is taken to support the claim that collective actions metaphysically reduce to individual actions. In this chapter I argue that this view is untenable. I raise principled doubts about its overall strategy, reconstruct an argument for it, and show with three different counterexamples that it fails.

The paraphrase argument expounds on the belief that statements about alleged collective actions are merely a shorthand for statements about individual actions. For example, "the Supreme Court strikes down the Defense of Marriage Act" is paraphrased as "the justices strike down the Defense of Marriage Act". While the former sentence is about a collective action — it suggests that a group of individuals does something — the latter sentence refers to the actions of individuals. Quinton (1975) presents the locus classicus for expressing the belief that statements about collective actions are merely a shorthand for statements about individual actions, which I call the *paraphrase hypothesis*.

> [A]ll statements about social objects are statements about individuals, their interests, attitudes, decisions and actions. But the predicates of these statements ... mention social objects in a way that is not practically or usefully eliminable, even if it is eliminable in principle (Quinton, 1975, 25).

Are sentences about collective actions merely a shorthand for sentences about individual actions? This paraphrase hypothesis is often expressed,[1] but it has never been fully defended.[2] The hypothesis is not obviously true; it demands a proof. Furthermore, it should be noted that the hypothesis makes a claim about "statements". How exactly does an observation about language bear out metaphysical facts about reduction? To answer this question, we need an argument.

With this chapter, I aim to rectify these shortcomings. I begin on a sceptical note in Section 2.1 and caution against undue optimism about the potential of paraphrase arguments in general. I then outline desiderata they should satisfy. In Section 2.2 I proceed more constructively and present a paraphrase procedure that systematically unpacks statements about collective actions into statements about individual actions. In Section 2.3 I reconstruct the argument for how this procedure vindicates the metaphysical reduction of collective actions. I then return to a sceptical outlook. I observe in Section 2.4 how this paraphrase procedure does not meet the desiderata discussed in Section 2.1. I show with three counterexamples different ways in which the paraphrase argument fails.

I need to make two important clarifications before I begin. First, this chapter focuses on the question of whether collective actions metaphysically reduce to individual actions. I take it to be a further question as to whether collective actions are eliminated by a reduction or whether they retain their status as entities (List and Pettit, 2011, 10; Sylvan, 2012). Eliminative reductionists deny that collective actions exist. Non-eliminative reductionists, in contrast, accept that there are collective actions even though they metaphysically reduce to individual actions. I am interested in the common ground between these

[1] Several authors refer to this hypothesis using the term "shorthand". French (1979, 211), Wall (2000, 195–96), Elster (2007, 13) and List and Pettit (2011, 2–3) are examples. Furthermore, some authors advance a paraphrase argument under a different heading. Watkins (1973), Quinton (1975, 23–25), Copp (1979, 178), Tuomela (1989) and Miller (2001, 10) are examples. Sawyer (2001, 563) attributes a "shorthand" view to Coleman (1994). A very similar hypothesis can be found in the debate on collective responsibility. See Cooper (1968) and Isaacs (2011, 82), who reject the analogue shorthand hypothesis for collective responsibility.

[2] To my knowledge only Ludwig (2007a,b, 2014a) and Massey (1976) can be read as attempting to show that the paraphrase hypothesis is true by giving rigorous accounts of how such a paraphrase might work in principle. Tuomela (1989, 476) acknowledges the need to show that the paraphrase hypothesis is in fact true. Short of doing so himself he suggests that a formal analysis of the logical form of collective action sentences should be used in a "deeper investigation". I read Ludwig (2007a) as undertaking this deeper investigation in the way suggested. We should not confuse the paraphrase hypothesis with a claim about constitution or grounding (see my comments below). The argument by Copp (1979) gives an account of constitution. And, unlike any paraphrase argument, he does not base his claims on an investigation of natural language.

two positions, which is the reductionist claim that collective actions *just are* the actions of individuals.

The second clarification is that this reductionist claim must be distinguished from the views of constitutionists. Constitutionists contend that collective actions are constituted by, or grounded in, the actions of individuals. But this claim is different from the reductionist claim. The constitutionist claim involves a grounding relation, whereas the reductionist claim involves an identity relation. The two relations differ importantly in their formal properties. The identity relation is symmetric, while the grounding relation is asymmetric.[3] Furthermore, the constitutionist claim does not entail the reductionist claim. A collective action may be constituted by individual actions and nevertheless be distinct, that is, non-identical, from them.[4]

Similarly, we must distinguish the reductionist claim from claims about contribution. The reductionist claim is about the agency of actions. It is concerned with whether the agent of an action is an individual or a collective. In contrast, the contribution claim is concerned with how an action comes about. The contribution claim is weaker than the reductionist claim; that is, contribution does not necessarily entail agency. An individual can contribute to a collective action without being an agent of it. It is consistent to claim, on the one hand, that any collective action is brought about by individual contributions, while contending on the other hand that there is a collective action distinct from the individual contributions. In this chapter, I leave questions of constitution, grounding, and contribution aside to focus solely on the reductionist claim.

2.1 Desiderata for a Paraphrase Procedure

A paraphrase is a relation between sentences. It relates one sentence, the shorthand, to another sentence, the longhand. A *paraphrase procedure* is a method for producing longhand sentences from shorthand sentences, such that the former paraphrase the latter. The procedure is employed to determine what exists. In most cases a "[p]araphrase is offered as an escape route from

[3]The identity relation is symmetric: when A just is B then it is also the case that B just is A. The constitution relation is asymmetric: when A is grounded in B then it is not the case that B is grounded in A.

[4]In fact, because the one involves an asymmetric and the other a symmetric relation, the constitutionist claim and the reductionist claim are incompatible. Nothing grounds itself, and likewise nothing constitutes itself.

excessive ontological commitments" (Jackson, 1980, 304). The procedure needs to meet several desiderata, particularly when it is used to play this theoretical role (von Solodkoff, 2014). I mention three desiderata and then focus on the last one.

First, a paraphrase procedure needs to be supplemented with an explanation of why the longhand sentence straightforwardly expresses ontological commitments, while the shorthand sentence does not (Alston, 1958). An eliminativist reduction rests on this asymmetry. An eliminativist about collective actions contends that the sentence "the justices strike down ..." straightforwardly expresses ontological commitments but its shorthand "the Supreme Court strikes down ..." does not. This asymmetry demands an explanation for two reasons. First, longhand sentences are not commonly used in language. Why should we rely on the longhand and not instead on the more commonly used shorthand to determine what exists? Second, a paraphrase usually merely puts in different terms what a shorthand sentence already expresses. What, in addition to restating the shorthand sentence in different terms, is a paraphrase procedure doing? This leads to a question about the logical relation between shorthand and longhand sentences.

Second, a paraphrase procedure should make the logical relation between shorthand and longhand sentences clear (Nolan, 2014b). Even though it is a necessary condition for an adequate paraphrase (which I formulate as *Minimal Adequacy* below), the relation cannot be the biconditional, that is, that the two sentences have the same truth value. Having the same truth value is insufficient to relate longhand sentences to shorthand sentences, because very many sentences have the same truth value but by no means stand in the paraphrase relation. A different proposal says that the relation between the sentences is a necessary conditional. That is, two sentences stand in the paraphrase relation if and only if, necessarily, when the shorthand sentence is true, the longhand sentence is true. Again, this is insufficient to relate longhand sentences to shorthand sentences. "Peter was more than 20 minutes late" necessarily implies "Peter was more than 10 minutes late", but the former is not a paraphrase of the latter. We can strengthen this further and take the logical relation to be the necessary biconditional. Under this proposal, shorthand sentences and longhand sentences necessarily have the same truth value. But this is still insufficient to account for the logical relation between longhand and shorthand sentences, because all necessary truths have the same truth value but by no means stand in the paraphrase relation.

The third desideratum is that the paraphrase procedure is universal. This means that it must produce an adequate longhand sentence for any shorthand sentence. What makes a longhand sentence adequate? I focus on the following necessary condition for adequacy.

> *Minimal Adequacy*.
> A longhand (individual action) sentence is adequate only if it has the same truth value as its shorthand (collective action) sentence.

If a paraphrase procedure fails to give an adequate longhand paraphrase for an alleged shorthand sentence, then the shorthand sentence might not actually be shorthand at all. A paraphrase procedure that fails to be universal also fails to support claims about metaphysical reduction. These claims need a universal paraphrase procedure because they are universally quantified, as for example the claim that *all* collective actions are just the actions of individuals.

Given these desiderata, the prospect of formulating a paraphrase procedure to support a metaphysical reduction seems unlikely (Nolan, 2014b). Indeed, I argue that in the case of collective actions, the best paraphrase procedure known so far is not universal. It therefore fails to support the reductionist claim.

2.2 Paraphrase Procedure

The paraphrase procedure I detail here rests on a regimentation of action sentences (Davidson, 1967; Ludwig, 2007a). Let an *action sentence* be any sentence in a natural language that refers to an action. Consider the following examples.

1. "I build a boat."
2. "We build a boat."
3. "The judge finds the defendant to be guilty."
4. "The judges find the defendant to be guilty."

The subject term is singular in sentences 1 and 3, and plural in sentences 2 and 4. Note that the form of the subject does not determine whether the sentence refers to individual or collective actions. Action sentences with a singular subject term can also refer to collective actions. An example is the action sentence "the Security Council condemned the violence in Syria".

Despite the singular subject term, this is a collective action insofar as the Security Council consists of several members. To avoid this complication, I focus on action sentences that, when they have singular subject terms, are about individual actions. These are *individual action sentences*. Analogously, all sentences with plural subject terms are *collective action sentences*.[5] Apart from differing in the form of their subject term, the sentences share a similar surface structure.

2.2.1 Individual Actions

Before developing a paraphrase procedure to unpack collective action sentences, it would be useful to understand how to regiment individual action sentences. Davidson (1967) developed the standard proposal to regiment action sentences in first-order logic. I indicate a regimented sentence with an R, such as $R(1)$ for the regimented form of sentence 1.

***R*(1).** $(\exists e)[\text{agent}(a, e) \,\&\, \text{building}(e) \,\&\, \text{of}(p, e)]$

The formula $R(1)$ reads: There exists an e such that a is an agent of e, this e is of the type building, and this e involves p. The letter e is a variable denoting an event. The letter a is an individual constant that refers to the person to which "I" refers in sentence 1. This person is an agent of the event e. In the formal language this is expressed with the *agent-event relation* "agent(a, e)". Any a that occurs in the first place of this relation is an *agent* of the event e. Any event e is a *collective action* if and only if its agent is a group consisting of at least two individuals. The expression "building(e)" is an *action predicate*, which says that this event is a certain action type. The individual constant "p" refers to the boat that is built. The boat is the patient of the action. The expression "of(p, e)" is a *patient-event relation*.

Consider sentence 3 as a further example: "The judge finds the defendant to be guilty". Let there be an event e where a judge a finds a defendant p guilty. The action predicate finding-guilty(e) expresses this action type. The action sentence 3 is then regimented as follows.

[5] There are two ways to understand action sentences with plural subject terms. They could be understood either *collectively*, in the sense that we are all building one boat together. Or they could be understood *distributively*, in the sense that each of us builds the boat independently of one another. Of these two options, only the former, the collective reading, is about collective actions. Since I am interested in collective actions, I focus on the collective reading. Tuomela (1989, 1995) cites an unpublished manuscript "Conjunction, Plurality, and Collective Particulars". from 1986 by R. Ware as the origin of this distinction. See also Ludwig (2007a, 361–62).

***R*(3).** $(\exists e)[\text{agent}(a, e) \,\&\, \text{finding-guilty}(e) \,\&\, \text{defendant}(p, e)]$

These examples illustrate a simple way of regimenting action sentences, which can be generalised with a schema. I abbreviate this schema as SI for *individual* action sentences. Sentences that are instances of this schema are *non-reductive regimentations*.

SI. $(\exists e)[\text{agent}(a, e) \,\&\, \textit{action-predicate}(e) \,\&\, \textit{patient-event relation}(p, e)]$

Regimenting action sentences in this way has limitations. It leaves aside aspects of a sentence that are not relevant to our discussion. For example, to keep the formal language simple we ignore the time at which an action occurs. Furthermore, we do not allow for different agent-event relations (ignoring so-called thematic relations or *Theta roles*). We admit different action predicates for various types of actions, such as building(e) or finding-guilty(e), but we hold fixed the relation agent(a, e). It is true that different action predicates may require particular senses of agency. But this distinction is only relevant for comparisons across sentences with different action types. In contrast, we investigate how individual action sentences relate to collective action sentences within one action type. So nothing speaks against holding the agent-event relation fixed. In fact, there are two important reasons in favour of doing so. First, in philosophy of action there is only one concept of being an agent of an action, that is, what is required of a to be an agent of e.[6] The agent-event relation encodes this concept of agency. Second, when attempting to reduce collective actions to individual actions we must keep fixed what we mean by "actions". Otherwise, we risk moving the bar of what is necessary for being an agent of an action depending on whether we talk about a collective or an individual agent. We keep the meaning of "actions" fixed across individual and collective actions, by allowing only one agent-event relation.

What does it take to be the agent of an action? Theories of intentional agency differ on this. But there are two widely accepted necessary conditions underlying the agent-event relation (Davidson, 1963). The first is an intentional condition, the second a causal condition. The intentional condition says that an a is an agent of e only if there is a description of e under which a intends e. This condition is not too demanding. It allows that I am an agent of actions

[6] Agency in philosophy must be distinguished from agency in linguistics. In philosophy of action, depending on the theory, agency is understood as a causal, moral, or rational concept, depending on the theory. In linguistics, agency is a semantic concept. For example, in the sentence "suspicion haunts the guilty mind", suspicion is an agent in this linguistic sense. Nothing of substantive interest for the present topic seems to depend on linguistic agency.

that turn out differently to how I intended. For example, when I try to wriggle a block out of a rickety tower in the *Jenga* game and the tower collapses, I am an agent of the collapse despite intending the tower not to collapse. The collapse can be described as an attempt to win the game, which is what I intended to do. The collapse is a side effect of my attempt, that is, something that I foresee might happen but that I intend not to happen.

The second condition for being an agent of an action is a causal condition. An a is an agent of e only if a stands in a causal relation to e. Different theories of action and causation give different interpretations of what this means. But we can bracket this issue for our present purposes. I will return to the conditions of the agent-event relation after presenting the paraphrase procedure for collective actions.

2.2.2 Collective Actions

How should we regiment collective action sentences? Our aim is to find a regimentation that can be used as a paraphrase procedure. For this purpose, we need a regimentation schema other than SI, which treats individual and collective action sentences alike. With SI, regimented collective action sentences would look just like regimented individual action sentences. Consider, for example, sentence 2: "We build a boat". Applying schema SI results in a formula that is orthographically identical to the regimentation of "I build a boat". The only difference is that the agent constant a refers not to an individual "I", but to a collective "we". Hence, it is a collective action. SI fails to unpack collective action sentences into individual action sentences.

Ludwig (2007a) develops a proposal that suggests an alternative regimentation schema.[7] The idea is to replace the plural subject term "we" with something like "each of us". We amend the earlier schema SI by adding "(Each x of a)" after the quantifier "$(\exists e)$". Furthermore, we replace occurrences of a with the new variable x.[8] This gives us a new schema. I abbreviate this schema as SC because we want to use it to regiment *collective* action sentences. Sentences that are instances of this schema are *reductive regimentations*.

SC. $(\exists e)$ (Each x of a)[agent(x, e) &
action-predicate(e) & *patient-event-relation*(p, e)]

[7] Ascribing this paraphrase procedure to Ludwig would be uncharitable. I merely adapt one among many proposals that Ludwig (2007a) introduces.

[8] More precisely, the quantifier (Each x of a) reads $(\forall x : x \in G)$ where G extensionally refers to a set of individuals.

In English, individual and collective action sentences may have the same surface structure. But proponents of the paraphrase argument contend that this similarity is misleading. Introducing the restricted quantifier (Each x of a) brings out an important difference between how collective and individual action sentences work. When we speak about a group doing something, we are in fact saying that individuals are doing something.[9] The new schema formalises what we really mean with a collective action sentence. It will help to illustrate this proposal with some examples. Consider the sentence "we build a boat" and its regimentation $R(2)$.

***R*(2).** $(\exists e)$ (Each x of a)[agent(x, e) & building(e) & of(p, e)]

The collective action sentence is regimented into a formula about individual actions. In English it could read as: "Each of us is an agent of an event that is a building of the boat". To many, this sounds wrong. $R(2)$ seems to say that each of us builds the boat. But each of us only contributes to the construction. Is there a flaw in the regimentation? No. We need to look at what $R(2)$ says precisely. It says that each of us *stands in the agent-event relation* to the construction of the boat. What does it mean to stand in the agent-event relation to an action? It could mean that an agent just contributes to an action in the right way. It does not mean that only one individual performs the respective action and no other.

The paraphrase procedure transforms any sentence in which we as a group stand in the agent-event relation into a sentence in which each of us individually stands in this relation. Is this paraphrase adequate? Presumably, as a group we stand in the agent-event relation to the construction of the boat. But is this also true for each of us individually? We can consider the two necessary conditions for the agent-event relation. On the intentional and the causal condition it may well be true that each of us stands in the agent-event relation to the construction of the boat. First, it seems plausible that each of us intends to build a boat. For each of us there is some description under which he or she intends to build the boat. Some of us want to hone their carpentry skills, while others simply want to get out of the house. Second, it seems plausible that each of us stands in a causal relation to the construction of the boat. Each of us contributes to the construction and if one slacks, then

[9] Von Solodkoff (2014) distinguishes between pragmatic and semantic versions of the paraphrase strategy. The title of the paper by Ludwig (2007a) suggests he is investigating a semantic phenomenon. However, many reductionists seem to argue instead for a pragmatic shift. When speakers utter collective action sentences, the proposition that they assert is not that of the sentence they utter (of the form SI); rather, they assert the proposition of a slightly different sentence (of the form SC).

his or her part will be left undone or will have to be built by others. So it is plausible to assume that each of us meets the intentional and causal conditions necessary to stand in the agent-event relation to the action in question. Hence, the regimentation $R(2)$ gives a minimally adequate paraphrase of what seems to be a sentence about a collective action.

Consider as a further example the case of a democratic election. Some argue that it presents a counterexample against the paraphrase procedure. I argue that the paraphrase procedure handles the case adequately.

5. "We elect Anne as the mayor."

***R*(5).** $(\exists e)$ (Each x of a)[agent(x, e) & electing(e) & of(p, e)]

According to $R(5)$, each of us stands in the agent-event relation to the election from which Anne emerges as the winner. Does each of us meet the two necessary conditions of the agent-event relation with respect to this event? Consider first the intentional condition. For each of us there is a description of the election under which he or she intends it. Some intend to elect Anne and some others intend to elect her opponent, while others only intend to fulfil their civic duties to vote. An agent can meet the intentional condition even if what results from the action differs from what he or she intends. Even for those who intend Anne's opponent to win, there is a description of the election under which they intend it.

Consider now the causal condition. Whether one stands in a causal relation to an event depends on the theory of causation and the set of necessary and sufficient conditions it stipulates. In some theories of causation, none of us individually stands in a causal relation to the event (cf. Lewis, 1973a). In such cases, $R(5)$ would not be an adequate longhand of sentence 5. The sentence would be false because the conjunct agent(x, e) would be false. In other theories of causation, each of us stands in a causal relation to the event (cf. Chockler and Halpern, 2004). As such, the regimentation $R(5)$ may be an adequate longhand in the sense that each of us meets the two necessary conditions of intention and causation. Even if these judgments might vary depending on the theory of causation, action sentences about democratic elections do not make a decisive case against this paraphrase procedure.

In summary, the regimentation schema SC defines a paraphrase procedure for simple collective action sentences. It produces longhand sentences that are minimally adequate. It provides a *linguistic reduction* of collective actions by giving an account of how sentences that seem to be about collective actions are

in fact just sentences about individuals doing something. However, further assumptions are needed to get from the existence of a paraphrase procedure to metaphysical conclusions. Presumably, such an argument follows Quine (1960). But I share the observation of von Solodkoff (2014, 571) that "whilst it's clear what role paraphrasing is supposed to play according to Quine, it is remarkably unclear how exactly this strategy works". Let us see if we can reconstruct how it works in the next section.

2.3 Reduction Argument

The paraphrase argument aims to show that collective actions metaphysically reduce to individual actions. It needs to show how the linguistic reduction of a paraphrase procedure supports a metaphysical reduction. I reconstruct this argument with two assumptions. The first assumption is that there is a universal paraphrase procedure that meets the minimal adequacy constraint as defined above.

> *Existence of a Paraphrase Procedure*.
> There is a paraphrase for any collective action sentence. This paraphrase has the same truth value as the corresponding collective action sentence.

In particular the assumption requires that the reductive regimentation of any collective action sentence has the same truth value as its non-reductive regimentation. It does not make a difference to a sentence's truth value whether it is regimented using schema SI or SC. But the reductionist claim that all collective actions just are actions of individuals does not yet follow. We require an additional assumption such as the following.

> *Paraphrase Implies Identity*.
> A collective action *just is* an individual action if and only if, given a true collective action sentence about this action, there is an adequate paraphrase of that action sentence into an individual action sentence.

This principle is the missing link between the existence of a paraphrase procedure and a metaphysical reduction. It states that the existence of a

paraphrase shows that collective actions just are the actions of individuals.[10] This assumption is reminiscent of Quine (1960, 241), who writes: "To paraphrase a sentence into the canonical notation of quantification is, first and foremost, to make its ontic content explicit." With this additional assumption, the reductionist conclusion follows.[11] The two assumptions form a valid argument for the conclusion that collective actions metaphysically reduce to individual actions. Is the argument sound? In the next section I argue that the first assumption is false. The paraphrase procedure fails to be universal and adequate.

2.4 Counterexamples

Reductionists tend to stipulate that there is an adequate and universal paraphrase procedure. Even if the procedure works for *some* collective action sentences, it does not work for *all*. In this section, I discuss three counterexamples in which the paraphrase procedure gives inadequate results. In these examples, the procedure leads to individual action sentences with a different truth value than the corresponding collective action sentences. Hence, the paraphrase procedure fails to be adequate and universal.

Each counterexample takes a different line of attack. The first is based on a central result of judgment aggregation. In this case, the intentional condition is true for the collective but not for the individuals. The second is based on a counterexample against functionalism. In this case, the causal condition is true for the collective but not for the individuals. The third counterexample explores the possibility that intentionality might be fundamental on the collective level. Since actions presuppose intentionality, reducing collective actions to individual actions is a non-starter.

[10] The assumption above is formulated for the specific domain of collective and individual actions. It is an instance of a more general domain-neutral principle.

> *Paraphrase Implies Identity (Domain-neutral).*
> An entity of type X *just is* an entity of type Y if and only if, given a true sentence about a token of type X, there is an adequate paraphrase of that sentence in terms of tokens of type Y.

[11] Eliminativists go a step further. They argue that only individual actions exist, and that there are no collective actions. For this conclusion they need a different assumption.

> *Paraphrase Implies Elimination.*
> An action exists only if, given a true action sentence about this action, there is no adequate paraphrase of that action sentence in terms of another action.

Formally, we can trace the failure in each of these cases to a common source: the collective meets the conditions to stand in the agent-event relation but the individuals do not.[12] The truth of an action sentence crucially depends on whether or not an agent stands in the agent-event relation to an event. Leaving the other conjuncts aside, a reductive regimentation (using schema SC) is equivalent to a non-reductive regimentation (using schema SI) if and only if the expression agent(x, e) in a reductive regimentation has the same truth value for every x in the domain of the (Each x of a) quantifier as the expression agent(a, e) in the non-reductive regimentation. *The paraphrase procedure fails because it is not a theorem that* (Each x of a) agent(x, e) *is a semantic consequence of* agent(a, e).

2.4.1 Discursive Dilemma

Consider again action sentence 4: "The judges find the defendant to be guilty". I first employ the paraphrase procedure to reductively regiment this sentence as an instance of schema SC. Compare this with a non-reductive regimentation under schema SI, with the constant a for "the judges".

$\boldsymbol{R_{SC}}$**(4).** $(\exists e)$ (Each x of a)[agent(x, e) & finding-guilty(e) & defendant(p, e)]

$\boldsymbol{R_{SI}}$**(4).** $(\exists e)$[agent(a, e) & finding-guilty(e) & defendant(p, e)]

For an adequate paraphrase, these two regimentations must be equivalent. However, there exist situations where the reductive regimentation (SC) is false while the non-reductive regimentation (SI) is true. Consider the example of a discursive dilemma (Kornhauser and Sager, 1993; List and Pettit, 2002, 2011; Tollefsen, 2002, 36; Nehring, 2005). Three judges have to decide on a civil case. The question before them is whether or not a defendant is guilty of a breach of contract. Assume there are three premises underlying the guilty verdict that are individually necessary and jointly sufficient. The defendant is guilty if and only if (1) there is a valid contract, (2) the defendant committed a breach of this contract, and (3) court papers were filed in accordance with all procedural requirements. Each judge intends to convict the defendant if and only if the defendant is guilty.

[12] The opposite can be the case as well when of several individuals each causes an event e and intends e to occur although each individual acts independently of one another. Yet because there is no true collective action sentence in such situations, these are not situations in which opponents of collective agency would have to apply a paraphrase procedure.

Table 2.1: Profile of the judges' opinions on the four propositions.

	valid contract	breach	due process	guilty
Judge 1	✓	✓	✗	✗
Judge 2	✓	✗	✓	✗
Judge 3	✗	✓	✓	✗
Collective	✓	✓	✓	✓

As Table 2.1 illustrates, the judges vote on each of the three conditions and each condition achieves a majority. Two out of three judges find that there was a valid contract, a breach of contract, and a procedurally sound filing. This implies that the defendant is guilty. But an apparent contradiction arises. There is no judge who finds the defendant guilty. If the judges were to vote on whether the defendant is guilty, they would agree that she is not guilty. But the rules of the court prescribe that a collective decision is reached by voting only on the premises. Hence, while none of the judges individually finds the defendant guilty, collectively they find the defendant guilty and the defendant is convicted.

The paraphrase $R_{SC}(4)$ says that each judge individually stands in the agent-event relation to the event of finding the defendant guilty. But this sentence is false for two reasons. First, a conviction is not an unintended side effect but rather it is something that is brought about intentionally. Yet none of the judges intends this event to occur, so none of them is an agent of the conviction. Hence, the paraphrase $R_{SC}(4)$ is false because the conjunct "agent(x, e)" is false.[13] Second, the conjunct "finding-guilty(e)" is false as well. "Finding the defendant guilty" describes an epistemic judgement. But no judge makes such a judgment. On the contrary, each judge finds that the defendant is *not* guilty and votes accordingly. These individual actions bring about the

[13]This argument rests on a contradiction between four premises. (1.) Each judge is an agent of the conviction. (2.) There is no other agent of the conviction. (3.) There is at least one agent of the conviction that also intends to convict the defendant. (4.) No individual judge intends to convict the defendant. Because this is a contradiction, one premise must be given up. I would suggest giving up the second premise and accepting that the court is a group agent that intentionally brings about the conviction; but opponents of group agency cannot do so. Note that a joint or shared action is not enough to deny the second premise because we require that the conviction is intentional. According to most accounts of joint action (cf. Bratman, 2014, 9; Searle, 2010, 60), there are only individual bearers of (plural) intentions and, according to the fourth premise, the individuals do not intend to convict the defendant. Denying either the third or the fourth premise comes at a considerable cost. First, when we argue that convictions are not intentional, we must accept that our justice system relies on events that no one intended. Second, when we argue that the judges individually intend to convict the defendant, we must explain how this intention coheres with their epistemic judgment that she is not guilty.

conviction, but it does not follow that the conviction is an action of the judges individually.[14]

Now consider the sentence $R_{SI}(4)$, on which there is a collective action. This sentence states that the judges collectively stand in the agent-event relation to the event of finding the defendant guilty. This non-reductive regimentation stays very close to the original sentence in English. So it is reasonable to suppose that if the original sentence is true, then this non-reductive regimentation should be true as well. But let me in addition sketch an elimination argument for the conclusion that the judges collectively meet the necessary conditions for the agent-event relation. Let us assume that finding the defendant guilty is an action and therefore requires an agent. Furthermore, there is at least one agent of this action that intends this action to occur. None of the judges individually meets this condition and therefore no individual judge is an intentional agent of finding the defendant guilty. By elimination, the judges collectively are an agent of finding the defendant guilty.[15] Hence, the collective action sentence is true while the paraphrase into an individual action sentence is false. This violates the minimal adequacy condition and the paraphrase procedure fails.

2.4.2 China's Avatar

I now advance a second counterexample against the reduction argument. This time we focus not on the intentional but on the causal condition. The group as a whole meets this condition, while the individual group members fail to do so. This counterexample adapts the China-body system of Block (2006). I call it China's avatar.

Suppose the members of the Chinese nation are arranged so as to form a functional duplicate of a brain. The structure in which they are arranged

[14]This view is compatible with several claims that have been defended about such situations. First, as Ludwig (2014a, 126) suggests, the conviction might be a mereological sum of the individual actions. But a whole need not be identical to each of its parts, hence the view that the conviction is non-identical to the individual actions is compatible with the claim that they stand in a mereological relation. Second, this view is compatible with the finding of Ludwig (2014a, 122) and Chant (2007) that, relative to what individuals intend, the consequences of individual actions may be unintentional. Third, it is compatible with a principle that, according to Chant (2007, 247), is widely accepted in philosophy of action, namely that "[d]escriptions of the form, 'bringing about E,' where E is a causal consequence of an action, refer to actions". The consequence of the individual actions may itself be an action when we accept that the court as a group is an agent of the conviction.

[15]We are assuming that either the judges individually or the judges collectively are an intentional agent of the conviction. See footnote 13.

and the dynamics in which they interact are identical to that of a brain. Each member simulates a brain neuron and uses a radio transmitter to connect her with others. All individuals have instructions on what to do in response to the signals they receive. Together they control a humanoid body, which instead of a brain has a remote control interface wired into its cranium. This body is China's avatar. The avatar and the members of the Chinese nation together form the China-body system.

I make two assumptions about the China-body system. First, the China-body system is a collective. Like many collectives, it may act as if it were one individual, but it really consists of many individuals. Second, the China-body system has intentions. The China-body system is a functional duplicate of a human individual. The two only differ in how these functions are realised. I take it for granted that differences in realisation do not make a difference concerning intentions.[16] Now suppose the China-body system intends to butter a slice of toast. In other words, it intends its avatar to move in such a way that its movements are an action of buttering a slice of toast. Sure enough, just as any individual would act on this intention, so does the China-body system.[17]

6. "The China-body system butters a slice of toast."

I contend that the paraphrase procedure fails for this sentence. The non-reductive regimentation is true but the reductive one is false.[18] This time, the problem is not the intentional condition but rather the causal condition underlying the agent-event relation. The China-body system as a whole stands in the right causal relation to the action of buttering the toast. But its members do not. Figuratively speaking, while it is true of the system as a whole that it butters the toast, it is not true of its parts.

There are at least three arguments for this conclusion. The first argument is that there are distinct causes. The collective, but not any individual, instantiates the relevant property that causes the slice of toast to be buttered. The second argument is that there are distinct effects. The event of buttering the toast is distinct from any individual action. The third argument is based

[16] It should be noted that I do not assume that the China-body system is conscious. I set the issue of phenomenal consciousness aside.

[17] To avoid an exception to the assumption made earlier that all action sentences with singular subject terms are individual action sentences, it could be rephrased as "The members of the Chinese nation butter a slice of toast."

[18] The two regimentations are:
$\boldsymbol{R}_{SC}$**(6)**: $(\exists e)$ (Each x of a)[agent(x, e) & buttering(e) & of(p, e)];
$\boldsymbol{R}_{SI}$**(6)**: $(\exists e)$[agent(a, e) & buttering(e) & of(p, e)].

on intuitions. It is intuitively false to claim that each individual member of the China-body system meets the causal condition for being an agent of buttering the toast.

The first argument contends that the relevant cause of the toast being buttered is a property that is only instantiated by the collective. This argument draws an analogy to the ontological status of the mental properties of individuals over and above the neurological properties of their brains (Kincaid, 1986). In the domain of individuals, we assume that the property of having an intention is distinct from properties that realise this intention on a neural level. The same intention might be realised by different configurations of properties on the neural level. In this way, only the intention, but not any property on the neural level, stands in a robust causal relation to the action. This robustness makes the intention the relevant cause explaining the action. By analogy, the intention of the Chinese nation is distinct from properties realising this intention on the level of its members. Analogous to the view about individuals' intentions, the same collective intention might be realised by different configurations of properties on the level of individuals. In this way, only the collective intention causes the action robustly and thereby stands in the relevant causal relation to the action of buttering the toast. This analogy is plausible because the Chinese nation forms a functional duplicate of a brain. It is one first argument to suggest that the paraphrase of sentence 6 is inadequate. Only the China-body system as a whole meets the causal condition for being an agent of buttering the toast, and hence only the collective stands in the agent-event relation to the action.

The second argument for the conclusion that the reductive regimentation is inadequate contends that there are distinct actions. On the one hand there is the collective's action of buttering the toast, while on the other hand there are the individuals' actions of operating their transmitters. We appeal to the individuation conditions of events to show that the collective action is distinct from any individual action. There are two ways to individuate events. First, according to a spatio-temporal individuation, the collective action and the individual actions are distinct events because they do not occur at the same locations and at the same time.[19] Second, according to a causal individuation, buttering the toast and responding to a signal are distinct because they adhere to different causal conditions. An individual intention is necessary and sufficient for the individual action to occur. However, no individual

[19] I do not claim that the collective action and the individual actions are disjointed. Each individual action is a strict spatio-temporal subset of the collective action.

operating her transmitter is necessary for the collective action to occur. If some individuals do not operate their transmitter correctly, others will compensate for their failure and the toast will still be buttered. Furthermore, an individual responding to a signal is not sufficient for the collective action of buttering the toast to occur. Therefore, the collective action is distinct from the individuals' actions because of differences in causal properties.

The final argument arises from intuition. It contends that buttering the toast is an action only on the part of the China-body system because intuitively, the individuals do not meet the causal condition for standing in the agent-event relation to buttering the toast. Collectively, the Chinese nation as a whole causes the action. But the individual members of the Chinese nation do not do so. The influence that each individual has on the occurrence of the action is vanishingly small (Braham and van Hees, 2009). Intuitively, individual causal responsibility decreases with increasing group size (Lagnado et al., 2013). Hence, taking our intuitions about causation as a guide, no individual member of the China-body system stands in the agent-event relation to buttering the toast. The paraphrase procedure fails to give an adequate paraphrase.

This second counterexample has similar ramifications to the first. The reduction argument is not sound because the paraphrase procedure fails to be adequate and universal. Some believe there is a better paraphrase procedure. But do we have evidence for this conjecture? In the absence of good evidence, it is reasonable to conclude that there is a collective action: the China-body system butters a slice of toast.

2.4.3 Hive Minds

Finally, we can question a tacit presumption of the previous cases to suggest yet another way in which the paraphrase procedure fails to be adequate and universal. The presumption on which all previous cases operate is that collective actions involve individual actions. But this is not necessarily true. There can be collective actions without individual actions. These corresponding collective action sentences cannot be paraphrased into individual action sentences. How are such cases possible? Actions require an intention. There may be a collective that has intentions and that can act but which consists of individuals who lack intentions. Without individual intentions there are no individual actions to which the collective action could be reduced.

To find collective actions that do not involve individual actions, we need to look outside the domain of human interactions. In human groups, collective intentions may be derivative in the sense that a human group has intentions in virtue of its members' intentions. But we can imagine a reversal of this relationship. In other species, collective intentions are the fundamental intentions from which individual intentions are derived, if there are individual intentions at all. These are cases of *hive minds*.[20]

Bees and ants seem to have fundamental collective intentions. An individual bee or ant is akin to a simple robot without intentions. Nevertheless, collective intentions may arise from the interactions of these robots. A bee hive can be thought of as "a kind of exposed brain that hangs quietly from a tree branch" (Seeley, 2010, 204). Similarly, Hofstadter (2000, 328) suggests that "ant colonies are no different from brains in many respects". The analogy is that just as intentions emerge from a brain that itself lacks intentions, intentions in a bee hive or an ant colony may emerge from individuals that themselves lack intentions. Similarly, Wilson (2001, S265) claims that "it is only groups of social insects, not individual members of those groups, that possess a mind ... a *group-only* trait".

Consider the following example by Watkins (1952, 188–89). When a bee hive splits and a part of the bee population is displaced from an existing hive, we can observe that the bees in the two resulting hives will change their behaviour so that in each new hive the same proportion of bees occupies each role as in the original hive. For example, one hive will start rearing a new queen to fill the vacant position. It seems that this reorganisation is an intentional action. It can be argued that a hive as a whole intends a certain internal organisation. There are at least two ways to make this argument. First, one can appeal to an instrumentalist theory of intentions and argue that ascribing the intention to maintain a certain proportion of roles to the bee hive is useful for predicting its behaviour (Dennett, 1989). Second, one can appeal to a functionalist theory of intentions and argue that the bee hive realises states that play the role of having an intention. Either way, a bee hive can be considered a collective agent that performs a collective action, even though no individual intentional actions are involved in this collective

[20]Conceivably, there are humanoid species that operate as a hive mind. For example, in the fictional universe of Star Trek there is an alien species called the Borg. They are cyborgs who are all connected in a network with a very high bandwidth. Under normal conditions, members of the hive only speak in the first person plural and have no concept of individuality.

action (cf. List and Pettit, 2006).[21] For a concrete action sentence, let us take the example of a bee hive choosing a new nest site.

7. "The bees decide to relocate to a new nest site at *L*."

In deciding between alternative nest sites, individual bees implement cognitive functions similar to the functions implemented by human cognitive systems (Seeley, 2010, 204). Similarly to the earlier comparison between the China-body system and a brain, this suggests an analogy between a bee hive and a human cognitive system. One can contend that just as a human agent has intentions that are distinct from his or her brain states, the collective intention of the hive is distinct from the states of individual bees.

This may sound similar to the counterexamples above, but the case of a hive mind shows something much stronger. What this counterexample suggests is that it is not possible to have a paraphrase procedure that is universal and adequate. In hive minds collective actions cannot be reduced to individual actions. There are no fundamental individual intentions to which the collective intention for an action could be reduced. Without individual intentions there cannot be an individual action. Any adequate paraphrase procedure will have at best a restricted domain that excludes the collective actions of hive minds.

Any paraphrase of sentence 7 into individual action sentences would be inadequate. No individual meets the intentional condition to stand in the agent-event relation. In contrast to the other counterexamples, a paraphrase in terms of individual actions is not available in principle. If hive minds are possible, then it is impossible to have a paraphrase procedure that is universal and adequate.

[21] This assumes that a bee hive is a collective. "Collective" can be understood in roughly three ways. (1) as an entity that consists of individual agents; (2) as an entity with interacting parts; (3) as an entity with interacting parts that are not continuously macroscopically connected and rearrange their position in apparently independent behaviour. The first understanding rules out hive minds by definition and thus begs the question of whether there are hive minds. The second understanding would consider brains to be collectives. Although this is not obviously unreasonable, it thus faces the threat of over generation. The third understanding classifies the bee hive as a collective but not the brain. For the purposes here, I adopt this third understanding of "collective".

2.5 Conclusion

In this chapter I have investigated whether collective actions metaphysically reduce to individual actions. I have constructed an argument that expounds on the paraphrase hypothesis that statements about collective actions are merely a shorthand for statements about individual actions. The argument assumes that the paraphrase hypothesis is true and that, therefore, collective actions are nothing but individual actions. Our evidence for the paraphrase hypothesis is the existence of the paraphrase procedure. But as we have seen with the counterexamples, contrary to what the reduction argument assumes, there is so far no adequate and universal paraphrase procedure. Moreover, the stringency of what is required of such a procedure renders its existence unlikely. In the absence of an adequate paraphrase procedure and setting aside other arguments, there seem to be collective actions over and above the actions of individuals.

My argument proceeded by first outlining desiderata for paraphrase procedures. I focussed on the minimal requirement that a paraphrase must not change the truth value of a sentence. Second, I put forward a procedure for collective action sentences and constructed the reduction argument, which claims that the existence of a paraphrase procedure is sufficient for showing that collective actions reduce to individual actions. Finally, I used three different counterexamples to demonstrate that the reduction argument does not succeed.

This chapter leads to two important suggestions. First, the failure of the reduction argument provides evidence in favour of two kinds of collective actions. There are *joint actions*, for which the paraphrase procedure works; these are just actions of individuals. But there are furthermore *corporate actions*, for which the paraphrase procedure fails; these are genuine collective actions (cf. Pettit and Schweikard, 2006). Second, the chapter shows that the paraphrase project is far from complete. Given the popularity of the hypothesis that collective action sentences are merely a shorthand for individual action sentences, we can see relatively little momentum in developing a procedure to prove this suspicion right. Until an improved paraphrase procedure is developed, the claim that collective action sentences are merely a shorthand for individual action sentences still awaits vindication.

Chapter 3

Agency and Embodiment

Many believe that groups are not real agents. Groups lack something that appears necessary for agency: a body. Without a body of its own, a group has to rely on others to act on its behalf. Statements such as "we never see groups acting; we only ever see individuals acting" commonly express this objection against group agency.[1] Their failure to qualify as agents undermines their prominence in moral philosophy. Groups are inappropriate objects of moral responsibility because group actions are, after all, just the actions of individuals. Furthermore without agency, groups cannot have rights or duties. In this chapter, I argue for a closer examination of groups' potential as agents and as objects of moral responsibility.

The issue rests on a larger question: to be an agent of an action, in what sense does someone need to have a body? The answer to this question affects not only the viability of group agency, but also theories of action. First, many theories of action analyse actions in terms of bodily movements. These theories must clarify precisely the sense of embodiment they presuppose. Second, embodiment matters for their application. How do these theories extend to agents in a virtual reality or to individuals augmented by artificial intelligence or robotics? Some contend that our agency is malleable and that "we human individuals, just are these shifting coalitions of tools" (Clark, 2004, 137). Others assign an essential, indispensable, or otherwise special role to

[1] Huebner (2014, 130) presents this objection in very similar terms. A locus classicus of this objection is Velasquez (1983, 6). For similar arguments see Margolis (1974, 254), May (1987, 113), Tuomela (1989, 471), Copp (1979, 178), Kutz (2000, 104), Ludwig (2007a, 376), and Ylikoski (2012, 32). Proponents of group agency consider this a problem as well (Pettit, 2007, 189). An early statement of this view is found in Kelsen (1949, 79): "When one speaks of actions and forbearances of a juristic person [i.e. a corporation], it must be actions and forbearances of human beings which are involved."

the natural human body (McDowell, 1996; Searle, 1980, 1984; Adams and Aizawa, 2011).

I contend that embodiment is not necessary for agency and argue against the Embodiment View.[2] This view consists of three assumptions. First, an agent needs to perform an action herself. Second, the performance of an action requires a bodily movement. Third, a bodily movement involves a movement of the agent's biological body. I show that each assumption can be relaxed. As each assumption is relaxed, the set of things that we consider to be actions broadens to include different kinds of disembodied actions. To distinguish them, I provide a taxonomy. The possibilities include proxy actions, extended actions, and extended movements. Actions of any of these disembodied kinds may be as relevant, and the agents performing them as much real agents, as any orthodoxly embodied human individual.

This chapter is structured as follows. After giving a minimal characterisation of agency in Section 3.1, I present the Embodiment View in Section 3.2. I then argue in Section 3.3 that the Embodiment View in some cases contradicts the minimal characterisation of agency. Since the Embodiment View consists of three assumptions, I investigate in Section 3.4 how any one of the assumptions can be given up to escape the contradiction. This gives us a taxonomy of disembodied actions.

3.1 Agency over Actions

The topic of this chapter is agency over actions, which is represented by the relation *a* is an agent of *x* that holds between an individual *a* and an action *x*. The agency relation has two properties that are relevant for the present argument. First, there must be at least one agent for each action. I call this first property *Coverage*; each action is covered by at least one agent. Second, only *intentional agents* stand in the agency relation. That is, any agent that stands in the agency relation has beliefs, desires, or intentions. I leave open what it is to believe, desire, or intend something. In this respect, all cases below are sufficiently clear.

[2] I use "embodied", "embodiment", and "embodied agency" interchangeably. It should be noted that the sense of embodiment, which is the topic of this chapter, is different from that of cognitive science (that cognitive functions are not restricted to the brain) and robotics (that design of robotic functions exploit body-world interaction loops). See Metzinger (2006) for a distinction. Furthermore, I use "collective agency" and "group agency" interchangeably as well.

Minimal Characterisation of Agency.
The relation a is an agent of x holds between individuals a and actions x, such that
Coverage. for each action x there exists an agent a, and
Intentional Agents. each a has intentions.

The Coverage assumption is plausible. Our intuitive understanding of agency requires that there cannot be an action without an agent. Likewise, the Intentional Agents assumption is plausible as well. In one way or another, actions must be related to an agent's mental states. No action can come from an agent who does not have any intentions at all. Of course, the two assumptions provide only a terse characterisation of agency and they fall far short of a conceptual analysis. Yet, between different competing analyses of agency, these two assumptions form common theoretical ground.

3.2 The Embodiment View

Take a moment to conjure up examples of agency. The actions that come to mind, such as pouring milk into your coffee, or shutting the door, are likely to involve bodily movements. One might be tempted to think that performing a bodily movement is necessary for being an agent of an action.[3] Yet there are cases that disprove this as a general claim. There are *mental actions*, such as making a decision, adding numbers, or imagining a tree, that do not involve bodily movements.[4] There are also *omissions*, such as not helping a friend or not attending a meeting. Most omissions do not involve bodily movements (Moore, 2010). However, virtually everyone believes that there is a distinct third kind of actions in addition to mental actions and omissions, namely, so-called *overt actions*. In contrast to mental actions and omissions, overt actions are taken to essentially involve the movement of a body (Mele, 2003, 5). What makes an action an overt action is precisely that it involves the movement of one's body. It is this specific kind of action for which the Embodiment View seems plausible.

I argue that the Embodiment View fails even for overt actions. The distinction between overt actions and other actions is not stable and on reflection, it

[3]The verb "move" is ambiguous and can be used either in a so-called *transitive* or an *intransitive* sense (Hornsby, 1980, 10). I return to this distinction below, suggesting that each respective sense of "move" relates to a different condition of the Embodiment View.

[4]This is, of course, not to say that nothing goes on, or moves, in the body of the agent of such mental actions.

collapses. While we may give a negative definition of overt actions along the lines of saying that they are those actions that are *not* mental actions, there is no notion of embodiment deserving of its name by which we can distinguish overt actions positively.

When analysing a concept, we often begin by examining paradigmatic instances of the analysandum, that is, of the thing we want to analyse. The paradigmatic instances of the concept of agency are actions that usually involve bodily movements. Hence, we can find many passages in the literature suggesting that agency is embodied. In what follows, we consider some of these passages to investigate how this necessary condition of embodiment may be understood.[5]

A prominent statement, which many take to express the view that agency is necessarily embodied, is found in Davidson (1971, 49).

> [A]ll primitive actions are bodily movements.

Davidson makes an identity claim. He contends not only that actions *require* movements, but that actions are *identical* to movements. Moreover, he states that the movements to which actions are identical are movements *of the body*. Smith (2012) defends the same idea.[6] Apparently realising that this would rule out mental actions and omissions, Davidson (1971) is quick to add that "bodily movement" should be interpreted "generously". In particular, "[t]he generosity must be openhanded enough to encompass such 'movements' as standing fast, and mental acts like deciding and computing".[7] However, this qualification is of little help to the problem I am going to raise. To respond to the cases I discuss, it is not "movement" that must be understood generously but the modification "bodily".

Many understand Davidson's expression "bodily movement" to mean the movement of not more than one human body. Otherwise, Davidson (1971) could not serve as the basis for arguments against the agency of groups, for which he is being used.

[5] To be clear, I do not claim that the interpretation I put forward of the literature below is fully charitable. Rather, I contend that the interpretation I give is not implausible.

[6] Smith (2012, 396) writes: "The standard story tells us that what the agent does, when he acts, is move his body in a certain way."

[7] Similarly, in order not to rule out omissions, Smith (2012, 389) is anxious to stress that "any way in which an agent might orient his body counts as a bodily movement for the purposes of the standard story".

> If, as Davidson has argued, all actions of individuals are their primitive actions ... it would follow ... that there are no primitive actions of collectives (Ludwig, 2007a, 376).

It should be noted furthermore that the argument against group agency assumes that an agent's body must have a particular kind of make-up, namely, that it is *biological*.

> Persons have (biological) bodies and perform bodily actions in contrast to collectives (Tuomela, 1989, 471).

Among the theories of agency that identify actions with movements of the human body, Davidson's is but one view on offer. A competing identity theory distinguishes between different kinds of bodily movements and contends that actions are identical to one particular kind of bodily movement.

> [A] physical action is a bodily movement, and *physical action* is a determinate of the determinable *bodily movement* (Haddock, 2005, 164) [emphasis in original].

This particular brand of identity theory is defended by Haddock (2005), McDowell (1996, 90), and Melden (1956, 523).[8] The idea is that there are on the one hand *mere movements*, which are *not* actions, and on the other hand *agential movements*, which *are* actions. An individual performs an action if and only if she performs a bodily movement of this particular second kind. It is clear from the context that "bodily" in "bodily movement" means a human body. We will also have to rely on context to clarify the meaning of the following passage from yet another theory of agency.

> [The concept of basic act-type] may not be restricted exclusively to "bodily-movement" properties, such as raising one's hand, turning one's head, lifting one's foot, etc. For the sake of simplicity, however, I shall henceforth assume that basic act-types include only bodily movement acts of this sort (Goldman, 1976, 68).

Goldman (1976) assumes that actions involve bodily movements only "for the sake of simplicity". But he does not indicate how it could be otherwise. It is not implausible to read him as contending that actions necessarily involve bodily movements.[9] And setting aside this particular example, there are

[8] In fact, Haddock explicitly builds his view on suggestions by McDowell. The two authors share the intuition that this identity claim is required because the agent would otherwise be alienated from her body.

[9] At least, he seems to contend that actions are embodied as a matter of technical necessity. We can understand the claim that actions involve bodily movements *necessarily* in either of two

ample passages in recent literature that can plausibly be read as claiming that agency is embodied.[10] Some authors make explicit what they mean by "body".

> [A]n agent is an entity that has a body and can make that body move in various ways. ... [O]nly creatures which have a biological origin are self-movers (Steward, 2012, 16–18).

In most cases, however, authors do not theorise explicitly what they mean by "body". It is not implausible to read them as individuating the body *narrowly*, that is, they understand "body" as one's biological make-up. In contrast, when we individuate the body *broadly*, an object counts as a body independently of its make-up. This alternative view understands "body" functionally and contends that whatever object is used for bodily functions is a body. Many philosophers subscribe to the functional individuation of the body. As we will see, doing so is one way to avoid the problem that I am going to describe.

But for the moment, we must not dismiss the view that the body is individuated narrowly. Many people deny that "body" should be understood functionally. Some deny this because they reject the functionalist proposal at large and see agency instead as a restricted and distinctively human phenomenon.[11] Others reject the particular kind of functionalism that underlies the broad individuation.[12] Moreover, as we will see later, while individuating the body broadly is a good response to some cases, other cases require a different response. In these cases, there is a place for individuating a body narrowly.

Between the passages above, we can see some overarching themes that come down to several key claims. I use these claims to define the Embodiment View as the following three necessary conditions.

---

senses. First, as a *metaphysical necessity*, that is, as what must be the case in all metaphysically possible worlds. Second, as a *technical necessity*, that is, as what must be the case in all possible worlds, which are accessible from the actual world using the limited effort and technology available at present. For example, flying to the moon was technically possible only by the mid 1960s but it has always been metaphysically possible.

[10] For example, Fischer and Ravizza (2000, 112) define their view in terms of bodily movements. See also Alvarez (2013, 107).

[11] In addition to Tuomela (1989) and Steward (2012), I think here of Taylor (1966), Searle (1980, 1984), Hornsby (2004), and Mayr (2011).

[12] An example for this group are Adams and Aizawa (2011). Generally, I think here of so-called *a posteriori* functionalists. See the related debate about the Extended Mind Hypotheses (Clark and Chalmers, 1998), and the question of what the "mark of the cognitive" is (Adams and Aizawa, 2011).

Embodiment View.

E1 a is an agent of x only if a performs x.

E2 a performs x only if a's body moves.

E3 a's body is individuated narrowly (as, for example, the biological body).

There is more than one way to understand the claim that agency requires a bodily movement. This is because the verb "move" is ambiguous and could refer to two different meanings (Hornsby, 1980, 5). Notice the syntactic difference between "*I* move my body" (two-place) and "my body moves" (one-place). In the former case, I move my body because I *make* my body move. This is the *transitive sense* of "move". In the latter case, my body may move because I am on a train, or because someone else moves it. This is the *intransitive sense* of "move".

The Embodiment View reflects both senses of movement. The first condition of the Embodiment View relates to "move" in the transitive sense.[13] The second condition of the Embodiment View relates to "move" in the intransitive sense. Contrary to what is sometimes assumed (cf. Alvarez and Hyman, 1998, 228), I will argue that overt actions do not even require bodily movements in the transitive sense.

The first assumption of the Embodiment View requires that an agent performs her actions herself. By "performs" we understand roughly mean "does" as it is intuitively understood in its colloquial sense. It rules out that a can be an agent of x when someone else performs x on a's behalf. The second assumption expresses the view that performing an action requires an agent's body to move. This is the condition by which many attempt to distinguish overt actions from mental actions. The third assumption specifies that by "body" we mean the biological body.

Paradigmatic cases of agency meet all three conditions. If I pour milk in my coffee then my biological body moves (E2 and E3 are true) and I perform the action myself (E1 is true). However, a theory of agency should also be able to handle other cases.

[13] This first condition does not involve the verb "move" but only "performs". Nevertheless E1 relates to the transitive sense of "move" insofar as many authors take this transitive sense of "move" to mean that an agent performs an action herself.

3.3 Agency Without Embodiment

Sometimes cases arise that pose a problem for the Embodiment View (cf. Moore, 2010, 32). In this section, I will use a counterexample to argue that the Embodiment View is not true for all actions, and that it is not even true for all overt actions. In particular, I argue that the Embodiment View contradicts the minimal characterisation of agency. Since we want to hold on to these basic assumptions about agency, the Embodiment View must be false. I develop the contradiction by amending a case that the Embodiment View initially is able to handle.

> **Twin Jim**. Twin Jim wants to break the window of an old garden shed. He picks up a stone, throws it, and the window breaks.

The case satisfies Coverage, because there is an agent (Twin Jim) for the action (breaking the window). It meets the Intentional Agents assumption, because Twin Jim is a young human adult without any cognitive impediments. Finally, the Embodiment View applies because Twin Jim has a biological body (E3) that moves to perform an action (E2), which Twin Jim does himself (E1). Hence, Twin Jim meets all of the above necessary conditions to be an agent of breaking the window.

Imagine a variation of the Twin Jim case. Suppose Jim can control a little robot just as he can control his arm. It should be noted that the required technology is already available; this is not science fiction (Carmena et al., 2003; Nicolelis and Lebedev, 2009; Ifft et al., 2013; Nair, 2013).

> **Jim**. Jim wants to break the window of an old garden shed. He has a ball-shooting machine that he can control with his mind. The machine fires and the window breaks.

The only difference between the cases of Twin Jim and Jim is how the event is brought about. Twin Jim uses a stone to break the window, and Jim uses a ball-shooting machine. A change in tools should not make a difference as to whether or not there is an action. According to the minimal characterisation of agency, Jim *is* an agent of breaking the window.[14] But according to the

[14]Coverage requires that if x is the action of the window breaking, there has to be a corresponding agent of this action. Either Jim or the machine is an agents of x. Because it is not an intentional agent, the machine is ruled out. Therefore, Jim is an agent of x.

Embodiment View, Jim is *not* an agent of breaking the window.[15] This is a contradiction. Let us consider the argument in its deductive form.

Argument Against the Embodiment View.
1. x is an action. (By symmetry to *Twin Jim*)
2. There must be an agent of x. (*Coverage*)
3. Either Jim or the machine is an agent of x. (Hypothesis)
4. The machine is not an agent of x. (*Intentional Agents*)
S. Therefore, Jim is an agent of x. (From 1. – 4.)
5. Jim's biological body does not move. (Hypothesis)
6. Jim is *not* an agent of x. (From 5. and *Embodiment View*)
C. A contradiction. *Embodiment View* is false. (From S. and 6.)

This argument is valid, but is it sound? Assumptions 2 and 4 are the minimal characterisation of agency, which we have taken for granted. But the remaining assumptions need to be defended against objections.

The first assumption is supported by symmetry considerations. Since breaking the window is an action in the case of Twin Jim, it must also be an action in the case of Jim. One might object that the two cases are different. Twin Jim breaks the window, but Jim does something else instead — for example, he fills the machine with stones — which in turn brings it about the window breaking. Breaking the window is an action in the case of Twin Jim, but not in the case of Jim. Hence assumption 1, which states that it is an action in both cases, is false.

This objection does not succeed. There is no argument for why breaking the window is not an action. The objection seems to assume that there is only one action. Filling stones in the machine and breaking the window cannot both be actions. But I see neither a compelling argument nor a clear intuition to support this assumption. Instead, it seems plausible that in the case of Jim breaking the window is an action.

The third assumption is that either Jim or the machine is an agent of x. One might object that this disjunction rules out a third alternative, namely that both together are an agent of breaking the window. It might be a joint action.

But there cannot be a joint action between Jim and the machine. Each participant in a joint action must have beliefs and intentions and the machine has

[15] Even though the ball-shooting machine moves, Jim's biological body does *not* move.

no such states. Hence, for the same reason for which it cannot be an agent, the ball-shooting machine cannot participate in a joint action.

The fifth assumption states that Jim's body does not move. Whether this is true is a matter of definition. According to the Embodiment View, "body" refers to the biological body (E3). By "movement" I understand a change in the macroscopic spatial region that the body occupies. Forming an intention or undergoing a change of brain states does not count as a movement. Hence, Jim's body does not move. Yet one might object that, in fact, Jim's body does move. He has to position the machine and fill it with stones. Contrary to the assumption, doing so involves movements.

However, Jim could position the machine and then decide not to break the window. There seem to be two actions here: making preparations and breaking the window. Jim's body moves when making preparations but it stands still when breaking the window. With respect to breaking the latter action, assumption 5 is true.

3.4 Three Kinds of Disembodied Actions

Agency is not necessarily embodied. In fact, the Embodiment View does not even apply to all overt actions. As we have seen, the case of Jim is a counterexample. The Embodiment View consists of three claims. To avoid contradicting the minimal characterisation of agency in cases like that of Jim, which of these three claims should we give up? I suggest we should give up all of them.

It turns out that the Embodiment View has concealed distinctions that are useful for our thinking about actions. Giving up each of its claims makes room for different kinds of disembodied actions that we may encounter now or in the near future. I call them proxy actions, extended actions, and extended movements. They correspond to relaxing the first, second, and third claims of the Embodiment View. Short of arguing that there actually exist instances of each kind of disembodied action, my aim is just to offer a taxonomy and to illustrate each kind of disembodied action.

Each kind of disembodied action is different. Specifically, they differ in which attributes they ascribe to whom (Table 3.1). First, who is an *agent* of the action? Second, who performs the *action*? Third, *whose body* moves in performing the

action? As one embodiment claim after another is given up, more and more attributes are ascribed to the agent.

Table 3.1: Overview of different kinds of disembodied actions; *a* stands for an individual and *b* for a tool or another individual.

	Agent	**Action**	**Movement**
(¬E1) *Proxy Action*	*a*	*b*	*b*
(¬E2) *Extended Action*	*a*	*a*	*b*
(¬E3) *Extended Movement*	*a*	*a*	*a*

3.4.1 Giving Up E1: Proxy Actions

The first kind of disembodied actions are *proxy actions* (cf. Ludwig, 2014b). We make room for this kind when we give up the claim E1 that an individual has to perform an action herself. There can be an agent *a* and an action *x* such that *a* is an agent of *x* despite not performing *x* herself. Instead, some other *b* performs *x* on *a*'s behalf.

The idea of proxy actions is by no means new. In *Leviathan*, Hobbes (1909 [1651], Ch.16) draws a distinction between the "author" and the "actor" of an action. He gives the example of an attorney representing a client in a court case. Recently, the idea has received renewed attention in the context of collective actions. Ludwig (2014b, 76) contends that an action is a proxy action if "one person ... doing something *counts as* or *constitutes* or is *recognized as* (tantamount to) another person or group's doing something". Similarly, Copp (1979, 177) defines "secondary actions" as "cases where persons may properly have actions attributed to them on the basis of actions of other persons". Against the backdrop of the Embodiment View, we see that what distinguishes proxy actions is that they violate claim E1. Hence we define proxy actions as follows.

> *Proxy Action*.
> Any *x* is a proxy action if and only if there is an *a* such that *a* is an agent of *x* but *a* does not perform *x*.

Several candidates for proxy actions come to mind. For example, a spokesperson performs proxy actions when delivering statements on behalf of the president (Ludwig, 2014b). The president might be an agent of delivering the statement without performing the action herself. Because E1 is given up,

performing an action oneself is not a necessary condition of being an agent of it. Delivering the statement is an action of which the president is an agent, but which is performed by the spokesperson.

For another example of a proxy action, consider the case of a bailiff performing an action on behalf of the court. Suppose that the court decides the Occupy camp is unlawful and intends that the camp be removed. It licenses a bailiff to remove the camp. In this case the bailiff performs a proxy action and removes the camp on behalf of the court or the state. What makes this a proxy action is that the court is an agent of the removal, even though it does not perform the action.

We now see how the failure of the Embodiment View makes room for the agency of groups. Groups never perform their actions themselves because they do not have bodies of their own. Any action of a group is therefore a proxy action. But since embodiment is not necessary for agency, the action of a group is an action all the same. We have seen that there is a difference between performing an action and being an agent of it. There can be actions, which individuals perform on behalf of a group but of which they are not an agent.[16]

3.4.2 Giving Up E2: Extended Actions

The second kind of disembodied actions are *extended actions*. We make room for this kind when we give up the claim E2 that a bodily movement is required to perform an action. There can be an agent a and an action x such that a is an agent of x and a performs x herself but a's body does not move in performing x. Instead, it is the body of another individual or tool that moves in the course of performing x.

> *Extended Action.*
> Any x is an extended action of an a if and only if a performs x but a's body does not move in performing x.

We find an example of an extended action in the case of Jim. He performs this action himself (E1 holds), and when his body is individuated narrowly (E3 holds), his body does not move (E2 fails).

[16] There might be two actions that are both performed by the representative. But the representative might be an agent only of one of the actions she performs; the principal would be the agent of the other action. I return to this in Section 4.2.2.

Extended actions are common in human–machine interactions. A machine may move as the agent performs the action while keeping perfectly still. Similarly to the case of Jim, an agent may control the movements of an artefact using a brain-machine interface (BMI) while keeping her own body stationary. With extended actions, human agency is no longer limited to its natural environment, but rather extends into the virtual domain (cf. Clark, 2004, 122). For example, when you use a BMI to control a mouse cursor on a screen or even your virtual self in a virtual reality, you perform extended actions. These are overt actions like any other overt action of which you are an agent and that you perform; the difference is just that your body does not move in performing these actions.

Another prevalent example of extended actions are remotely controlled military drones. In contrast to the neural inputs required by a BMI, drones are controlled via a remote pilot's manual inputs. Suppose the pilot can control the drone's movements by programming tasks such as flying a patrolling pattern or reacting to unexpected ground activity. These movements of the drone are actions that the pilot performs. Similar examples of this sort include driving the Mars exploration rover or, given their high degree of automation, piloting a modern commercial aeroplane (Clark, 2004, 25).

3.4.3 Giving Up E3: Extended Movements

The third kind of disembodied actions are *extended movements*. We make room for this kind of action when we give up the claim E3 that the body is individuated narrowly as the biological body. There can be an agent of an action x who performs x herself and her body moves. This is an overt action but the agent's biological body remains stationary. The agent's body moves nevertheless because there is a distinct object b, which is individuated as part of the agent's body, and it is only this b-part of the agent's body that moves when the agent performs the action.

> *Extended Movement*.
> A movement of a is extended if and only if there is an action x such that a's body moves in performing x but a's narrowly individuated body does not move.

Applied to the case of Jim, this escape route implies that the machine is a part of Jim's body. This is implausible because the ball-shooting machine

is not connected with Jim's biological body in the right way. But perhaps parts of a body do not need to be connected. Instead, one could contend that the body is just those objects under an agent's control, regardless of how the agent is connected to them.

> If movements [that are the first perceived effect of the operation of a mental cause] occurred in another body, or inanimate objects, then our concept of what constituted our body would tend to expand. For it is part of the concept of our body that it is that which is under the immediate control of our will (Armstrong, 1968, 146).

One common class of extended movements is when artefacts augment the biological body. A prevalent case is the use of prosthetic limbs. Think, for example, of the sprinter Oscar Pistorius or of Captain Hook. The former's prosthetic legs and the latter's hook are individuated as parts of their bodies.[17] If a prosthetic is permanently attached to an agent's biological body and if its use is transparent to the agent (moving the prosthetic is something the agent can just do), then it is plausible to think that it forms a part of the agent's body. To the agent it may even feel like the limb is a part of her body (Botvinick and Cohen, 1998).

In another class of extended movements, an artefact not just augments but also temporarily replaces an agent's body.

> **Avatar**. Jake is a paraplegic. There is an artificial humanoid body into which he — and only he — can log in. This body is his avatar. During the time in which he is logged in, Jake perceives everything as if he were the avatar. He controls the avatar's movements and has no conscious awareness of his biological body.

It could be argued that Jake has two bodies, his biological body and his avatar body, which he controls at various times. When the Avatar moves, these are movements of Jake's body. Already with existing virtual reality technology, agents may indeed feel like they actually are the avatar that they control (Ehrsson, 2007; Lenggenhager et al., 2007; Pomes and Slater, 2013). Of course, these cases raise various issues about consciousness and personal identity (cf. Dennett, 1981). But setting these aside, the Avatar case is another example of an extended movement.

[17] While most current prostheses are still controlled by movements of the biological body, we can imagine prostheses that are controlled by a BMI or signals from the nervous system.

3.5 Conclusion

In this chapter I have argued that agency does not need to be embodied. I have broadened the domain of agency to include disembodied actions, among which I have distinguished three different kinds (Table 3.2). First, others can act on behalf of an agent. Individuals perform actions on behalf of a group agent. Second, the movements of others can form the actions of an agent. The patrol pattern that the drone flies autonomously is an action of the pilot who controls it remotely. Third, the movements of others can be movements of an agent herself. An artificial or virtual extension such as a prosthesis can qualify as part of an agent's body. This taxonomy results from relaxing each of the three claims underlying the Embodiment View.

Table 3.2: Overview of different examples of disembodied actions.

Kind	**Examples**
Proxy Action	*Group actions:* Spokesperson, Bailiff
Extended Action	*Human–machine interactions:* BMIs, Drones
Extended Movement	*Augmentation:* Prostheses, Avatar

In the first half of this chapter I have developed the Embodiment View and argued that it does not even hold true for so-called overt actions (which contrast with mental actions). I have suggested that many authors can be understood as holding this view. But the Embodiment View is relevant not because of who holds it, but because of the arguments that depend on it.

Most prominently, the Embodiment View rules out the agency of groups. If agents must be embodied, then groups are not real agents. By arguing that agents do not need to be embodied, I have defended the possibility of group agency. We have seen that group actions are a specific kind of disembodied actions. They are proxy actions, which individuals perform on behalf of the group. Despite lacking a body and depending on others to act on their behalf, groups may be agents.

Part II

Locating Agency

Chapter 4

Responsibility in Hierarchical Groups

In this chapter I put forward a novel theory of agency in hierarchical groups. Examples of such groups include bureaucratic organisations, private corporations, and the military, to name a few. While jurisprudence has developed a significant amount of literature on this topic, philosophical accounts of group agency have largely overlooked the problem of explaining who can be held responsible for the actions of hierarchical groups.[1] Criminal law encompasses many cases with commanders on trial. Think of the Nuremberg trials, the Eichmann trial, or the killings of fugitives at the Berlin Wall.[2] The prosecution requires a justification for how a commander can be held

[1] The large body of literature on collective responsibility addresses a different problem. Collective responsibility arises usually from joint actions or from horizontal groups *without* a hierarchy. It describes situations in which a *group as a whole* is held responsible. See, for example, French (1979), May (1987, 1996), Pettit (2007), List and Pettit (2011), and Isaacs (2011). In contrast, in hierarchical groups *individuals* are held responsible for actions performed by others. While the topic has been largely overlooked in philosophy (but see Wasserstrom, 1980, Walzer, 2004, and Shapiro, 2014), it is widely discussed in international criminal law. See Roxin (1963, 2011) and, for example, Smidt (2000), Danner and Martinez (2005), Weigend (2011), Eldar (2013), and DeFalco (2013). Similarly, Feinberg (1970b) draws a distinction between collective responsibility on the one hand, and responsibility in hierarchical groups on the other.

[2] The most prominent case for the development of legal doctrine was that of Tomoyuki Yamashita (Smidt, 2000; Danner and Martinez, 2005). More recently the International Criminal Court (ICC) developed a similar doctrine to ascribe responsibility for crimes committed in the context of hierarchical groups. See ICC (2007, 2008), Weigend (2011), and Eldar (2013). Furthermore, the question of agency and responsibility in hierarchical groups arises in tort law.

responsible for the actions of her soldiers. In this chapter I develop such a justification from the vantage point of philosophy of action.[3]

I present the investigation using the case of autonomous drones. Advances in technology lead to moral quandaries involving machines making life-or-death decisions. The question of who can be held responsible for the actions of autonomous drones is considered to be a major challenge in the ethics of artificial intelligence (Sparrow, 2007; Cordeschi, 2013). Allegedly, a so-called responsibility gap arises when future drones will be, on the one hand, advanced enough to form intentions and make decisions on their own but, on the other hand, not advanced enough to be held responsible for them. I take this problem as a case study for hierarchical groups generally. Consider the following case.[4]

> **Drone Deployment**. A future autonomous military drone bombs a column of enemy soldiers who had indicated their desire to surrender. A commander had ordered the drone to patrol the region and engage legitimate targets. The drone wrongly identified the surrendering soldiers as legitimate targets.

Who can be held responsible for the bombing? Many argue that no single entity within the hierarchy can be held fully morally responsible for the bombing (Sparrow, 2007; Matthias, 2004).[5] If this is true, we face a responsibility gap.

> *Responsibility Gap*.
> An event x gives rise to a responsibility gap if and only if
> 1. no individual agent can be held fully morally responsible for x;
> 2. but had x been the action of a human person, then she could be held fully morally responsible for it.

[3] It should be noted that the investigation here is importantly different from many investigations in philosophy of action. This chapter does not engage in a conceptual analysis of intentional action and I am not committed to any particular ontology of actions. This is a work of philosophy of action insofar as it suggests a sufficient condition for being an agent of an action, in the sense of agency that is necessary for responsibility (cf. Baker, 2000, 149; Pettit, 2007, 175). I use this condition to develop a theory of agency in hierarchical groups.

[4] This case differs from that discussed by Sparrow (2007), who assumes that the drone deliberately bombs illegitimate targets and does not make a mistake. The present case is simpler without affecting the generality of the discussion.

[5] Furthermore, Steinhoff (2013, 180), while disagreeing with Sparrow (2007) overall, accepts his premise that "[t]here are many 'acts' an automated weapon system might commit, including instances of the killing of innocent people, for which neither the programmer of the weapon, nor the commanding officer, nor the machine itself can (justly) be held responsible".

Had the commander deployed a human soldier instead of a drone, we would be quick to hold this soldier responsible for her mistake in bombing the soldiers who had surrendered. A drone, in contrast, is an inappropriate target for blame. Who can be held responsible? There have been numerous suggestions about how to close the alleged responsibility gap.[6] I argue that the commander can be held responsible for the bombing because the drone acts on her orders as part of a hierarchical group. In doing so, I offer a theory of agency in hierarchical groups of which the military is but one example. This theory vindicates advice that the United States army gives to its personnel: "Commanders are responsible for everything their command does or fails to do."[7] Furthermore, the theory supports and sheds light on different legal doctrines that have been developed to tackle the question of agency and responsibility in hierarchical groups.[8]

A theory of agency in hierarchical groups must address two challenges. One challenge is to explain how a superior, such as the commander, can be held responsible for the actions of subordinate, such as the drone. Another challenge is to explain how a superior's responsibility can be greater than that of a subordinate. Although it cannot be seen in the case of drones, it is a familiar phenomenon that responsibility increases with hierarchy. The Jerusalem District Court put it as follows during the Eichmann Trial (1961, 197).

> [T]he extent to which any one of the many criminals were close to, or remote from, the actual killer of the victim, means nothing as far as the measure of his responsibility is concerned. On the contrary, in general, the degree of responsibility increases as we draw further away from the man who uses the fatal instrument with his own hands and reach the higher ranks of command.

Such *increasing responsibility* is a characteristic phenomenon of hierarchical groups (Lawson, 2011, 240). Despite the prevalence of hierarchical groups

[6] Lin et al. (2008) suggest holding the commander responsible but provide no argument for this solution. Lokhorst and van den Hoven (2011) suggest that the programmers of the drone can be held responsible. Schulzke (2013) sees a shared responsibility between agents in the military hierarchy and the developers. Steinhoff (2013) contends that neither the commander nor the programmers can be held responsible, but rather it is the politicians who should be held responsible. Pagallo (2011, 353) suggests that solving the problem requires "a new kind of ... responsibility".

[7] See US Department of the Army (2014, 2-1b).

[8] There are two main families of doctrines. One is the command responsibility doctrine developed mainly in international case law beginning with the Yamashita trial (Smidt, 2000; Danner and Martinez, 2005). Another is the doctrine of indirect co-perpetration developed originally in German criminal law (Weigend, 2011).

there is no rigorous philosophical explanation for it. In this chapter I aim to fill the gap by giving an account of how the commander has control over the bombing. Being in control of an outcome is sufficient for potential responsibility.[9] With this account of control we not only avoid the responsibility gap but can also understand more generally how agency and responsibility work in hierarchical groups and how responsibility increases with hierarchy.

This control-based account gives rise to a further challenge. How does the commander control the bombing, given that the drone performs it autonomously? We need to avoid an apparent contradiction. When someone performs an action autonomously it implies that she has control over it. This seems to rule out that there is someone else who has control over the same action. Is there any plausibility to the view that a commander controls the actions of her soldiers, who themselves in turn have control over their own actions as well?

I show that there is no contradiction between the commander's control and the drone's control over its decision to bomb the soldiers. They each control events on different levels. There is a sense of control in which an agent can be held responsible but which leaves room for the autonomy of others. In developing this thought, I draw on a proposal that has recently been advanced in discussions on mental causation.[10] I use it to put forward a novel theory of agency in hierarchical groups.

This theory has an important implication. In strictly hierarchical groups, there is no collective responsibility. There is nothing for which the group as a whole is responsible; all responsibility is individual responsibility.[11] In an ideal pyramid-like hierarchy, the agency of the individual at the top

[9] Similarly Roxin (2011, 200) states "Loss of proximity to the act is compensated by an increasing degree of organizational control by the leadership positions in the apparatus." This doctrine is adopted by the ICC. "[P]rincipals to a crime are not limited to those who physically carry out the objective elements of the offence, but also include those who, in spite of being removed from the scene of crime, control or mastermind its commission because they decide whether and how the offence will be committed" (ICC, 2007, §330). See also ICC (2008, §§484–86). For a comparison of this doctrine with rival approaches see Danner and Martinez (2005), and Manacorda and Meloni (2011).

[10] In particular, I follow List and Menzies' (2009) proposal of difference-making. An alternative account could be spelled out in the programming model of Jackson and Pettit (1990).

[11] As we will see, in a strictly hierarchical group there are no collective actions. By "strictly hierarchical group", I understand a hierarchical group (as defined below) in which only individuals stand in the authority relation such that there is exactly one individual who is *at the top*, that is, she is connected to all other group members via a path in the authority relation. This contrasts with horizontal groups, in which all individuals make a decision together. To the extent that we find authority there at all, in this latter type of group the authority relation is symmetric. Each type is an idealisation; most groups have elements of both types.

extends through all the lower ranks and carries with it her responsibility for the outcomes that eventually ensue.

This chapter has three parts. In Section 4.1 I give an argument for how a responsibility gap arises. In Section 4.2 I put forward a novel theory of agency in hierarchical groups based on recent advances in theories of causation. In Section 4.3 I apply this theory to respond to the responsibility gap problem in the case of future autonomous military drones.

4.1 The Responsibility Gap and Related Issues

First, I need to clarify the notion of a responsibility gap. Second, I identify how the responsibility gap for autonomous drones arises. Third, I distinguish the issue that gives rise to the responsibility gap (the agency–responsibility issue) from two related issues (the individual–collective issue, and the success–failure issue).

4.1.1 The Conditions for a Responsibility Gap

A situation gives rise to a responsibility gap if it meets two conditions. First, there exists an event for which nobody can be held fully morally responsible.[12] Second, if a person had been involved, then there would have been someone who could have been held fully morally responsible. Specifically, focussing on the case of drones, if a person had been involved instead of a drone, then someone would have been fully morally responsible. The first condition is the *responsibility void* condition.[13] The second condition is the *because of drones* condition. A situation involving a drone gives rise to a responsibility gap of the kind that we are focussing on here if and only if both conditions are met.

Let me clarify three points. First, in this chapter we examine *future drones*, which are not remotely controlled by drone pilots. Rather, these future drones are autonomous in the sense that they are given a mission and they make life-or-death decisions in carrying out this mission just as human soldiers

[12] This, like the definition above, defines a *thin responsibility gap*. Not all responsibility gaps of this kind are morally problematic. Think of a landslide that was caused by heavy rain. This event gives rise to a responsibility gap in the present formulation. This is because if the landslide had been the action of a human person, then this person could be held fully morally responsible for it. We can define a *thick responsibility gap* by narrowing the domain of x from events to actions.

[13] I borrow this term from Braham and van Hees (2011).

would. When faced with a situation such as that described in the Drone Deployment case, they do not defer to a human to decide whether to engage the targets. No human is involved in carrying out the mission.

Second, the property that we want to locate with an individual has to do with *moral responsibility*. In particular, the issue here is fitness to be held responsible in the sense of backward-looking moral responsibility.[14] The question is not about forward-looking responsibility in the sense that an agent is under an obligation concerning her future conduct. Instead, the question is whether after the fact there is an agent who is a liable target for blame or sanctions with respect to an action or state of affairs.[15] Furthermore, holding an agent responsible should be distinguished from accountability in its legal sense. I might be accountable for the misbehaviour of my children but I am not morally responsible for what they do (cf. Pettit, 2007, 173). Or conversely, when I could easily help a person in dire need but refrain from doing so, I am morally responsible but not legally accountable.

Third, the definition requires that an agent is *fully responsible*. This requirement rules out that the commander has only partial responsibility for the bombing.[16] If the commander had only partial or shared responsibility, there would still be no agent who is individually fully responsible.[17] Holding the commander partially responsible would hence not be good enough to avoid the responsibility gap.[18]

[14]For simplicity, while I mean "an agent is fit to be held responsible", I write "an agent is responsible" or "an agent can be held responsible".

[15]This is one way in which this investigation differs from, for example, Walzer (2004).

[16]For example, Schulzke (2013) defends the view that the commander and the developers of the drone together share responsibility for the bombing.

[17]Of course, shared responsibility and partial responsibility might refer to different phenomena. We shall not be satisfied with a proposal that fails to go beyond *either* of these two.

[18]We have good reason to require full individual responsibility. To begin with, it is not clear what shared or partial responsibility is exactly, and what it would permit us to do to those who have it. Furthermore, shared responsibility suggests that responsibility is distributed equally among individuals. This would neglect the fact that individuals contribute differently. In hierarchical groups, some individuals have more power than others and this should be reflected in the degree to which they are held responsible. I resist shared responsibility only insofar as it does not clarify the size of an individual's share. The proposal I put forth here offers ways to think about shared responsibility. For example, you could take an individual's shared responsibility as the fraction of her degree of agency over her superior's degree of agency — both of these notions will be clarified below.

4.1.2 Agency–Responsibility Issue: Drones

The alleged responsibility gap arises because of a gap between agency and responsibility. The conditions for *being an agent of an action* are less demanding than the conditions for *fitness to be held responsible for an action*. Suppose we hold you responsible for a bank robbery. Unless we got the wrong person, this entails that you must have had something to do with the robbery in the right sense. I suggest that we understand the sense of *having something to do with* the robbery that is relevant for responsibility as *being an agent of* the robbery. I take it for granted that responsibility requires agency.[19] We keep the issue of how agency is defined as a task to return to later.

> *Responsibility Requires Agency*.
> For any a and any x, if a is responsible for x then a is an agent of x.

The reverse requirement does not hold. Agency does not require responsibility. There are additional necessary conditions for being fit to be held responsible for an action above and beyond the conditions for being the agent of an action. To be the agent of an action is among the necessary conditions for responsibility. But responsibility may also require that the agent have access to relevant information.[20] Or, more vaguely, responsibility may require that the agent possess moral concepts or that she understand that she faces a decision where something of moral relevance is at stake. Since they are surrounded by deep controversy, I leave open what exactly the necessary conditions for responsibility are. I only claim that the conditions for responsibility are more demanding than the conditions for agency. This is the *agency–responsibility issue*.

This space between agency and responsibility is where future autonomous drones are located. Future autonomous drones might be advanced enough to meet the conditions for being the agent of an action but they might not be

[19] This basic assumption is uncontroversial when we understand agency broadly enough. As the discussion in this chapter suggests, there are different ways to be an agent of an action. In particular, it is not necessary to perform an action oneself to be an agent of it. To my knowledge there are only two attempts to reject the notion that responsibility requires agency. The first rests on specific assumptions about the metaphysics of omissions (Sartorio, 2004). The second assumes a noticeably strong notion of agency (Dempsey, 2013). See also my discussion of this assumption in Section 1.3.

[20] Compare, for example, the three conditions for responsibility put forward by Pettit (2007). See also Feinberg (1970a).

advanced enough to meet the conditions for fitness to be held responsible for that action.[21] A responsibility gap may arise.

The first condition for a responsibility gap would be met when nobody can be held responsible for the bombing. Suppose that there is only the drone and the commander and nobody else. Neither of them can be held responsible for the bombing. The commander cannot be held responsible because she is *not* the agent of it. Sparrow (2007, 71) writes: "If the machines are really choosing their own targets then we cannot hold the Commanding Officer responsible for the deaths that ensue."[22] However, the drone cannot be held responsible for the bombing either. While the drone might be an agent of the bombing, the drone does not meet one of the conditions necessary for responsibility.[23] The drone might lack a proper understanding of the moral concepts or it might lack access to the relevant evidence.[24] We face a responsibility void. In a deductive form the argument is this.

Responsibility Void Argument.

1. Only the drone or the commander is fit to be held responsible for the bombing.

2. The commander is not an agent of the bombing.

3. Responsibility requires agency.

S_1. The commander cannot be held responsible for the bombing.

4. The drone does not meet a condition necessary to be held responsible for the bombing.

S_2. The drone cannot be held responsible for the bombing.

C. Nobody can be held responsible for the bombing.

[21] Consider an analogy to the case of group agency. An unincorporated group meets the conditions to count as the agent of an action, but it cannot be held responsible for this action (Pettit, 2007).

[22] Note that the argument of Sparrow (2007, 71) implicitly makes further crucial assumptions, namely, that there is only *one* action (the bombing) and that there is at most one agent for each action (because the drone is an agent of the bombing, the commander cannot be an agent of it). But this argument is not sound. First, on the account I develop below, there are *two* actions. Second, it is plausible that there are things we do together. Such joint actions are actions with more than one agent.

[23] For example, Sparrow (2007, 72) contends that responsibility requires punishability "to satisfy our psychological need for revenge". This means that an agent can be held responsible only if the agent can be subjected to suffering "of the sort that we find morally compelling".

[24] There are different arguments for the subconclusion that the drone cannot be held responsible for the bombing because it fails to meet this or that condition, depending on how the necessary conditions are spelled out. To avoid getting sidetracked into a full analysis of responsibility that I want to set aside here, I take it for granted that there is at least one condition necessary for holding an agent responsible that the drone fails to fulfil.

The second condition for a responsibility gap is also met. Had a person instead of a drone performed the action, the person could be held responsible. In sum, we have a valid argument for a responsibility gap. If this argument is sound, then a drone deployment would give rise to a responsibility gap. Should we be worried?

I reject assumption 2. I argue that the commander *is* an agent of the bombing. Assumption 2 is false because the drone operated on the commander's orders and there is a sense in which the bombing is under the commander's control. If an individual controls whether or not an action occurs, then this individual counts as an agent of the action regardless of whether the agent performs this action herself. In hierarchical groups, where some wield power and authority over others who act under their command, agency and responsibility thereby traverse along hierarchical structures of command. No responsibility gap arises in hierarchical groups, or so I will argue. I spell out this view in detail in the rest of this chapter.

While some have suggested that the commander is an agent of the bombing and should be held responsible (Lin et al., 2008), others have dismissed the idea that the commander has control over the bombing (Sparrow, 2007, 72). This divergence calls for a clearer understanding of the characteristics of hierarchical groups, and for an account of how individuals exercise agency on their level of the hierarchy within those bounds and while following the instructions imposed on them from above. The key to understanding hierarchical groups is an adequate notion of control. In this chapter I put forward one way of understanding control that answers the main questions about hierarchal groups.

Before developing this account, to get a clearer view we need to distinguish two related issues. The first issue concerns collective responsibility. The fact that a decision is made collectively might stand against holding any individual member of a collective responsible for the results of a collective decision. This is the *individual–collective issue*. The second issue concerns intentions. The fact that an agent did not intend the results of her action might count against holding the agent responsible for them. This is the *success–failure issue*. These issues are red herrings. They pull towards different problems that do not genuinely have anything to do with hierarchical groups or future autonomous drones. It is nevertheless worth discussing these issues for three reasons. First, the literature has appealed to each of these considerations to suggest that a commander cannot be held responsible for

the bombing. Second, discussing these related issues helps to characterise the distinct topic of hierarchical groups more clearly. Third, an important upshot of discussing the success–failure issue will be relevant later on in the chapter when I define risky actions and argue that responsibility does not require success.

4.1.3 Individual–Collective Issue: Horizontal Groups

In some situations, a collective as a whole can be held responsible, while no individual member has full responsibility.[25] For example, in March 1987 the *Herald of Free Enterprise* capsized when leaving the Belgian harbour of Zeebrugge, killing almost 200 people. While nobody was convicted, an inquiry found that the shipping company as a whole could have been held responsible for negligence (Colvin, 1995, 16). This is an instance of the *individual–collective issue*. Some suggest that this issue is relevant for the case of autonomous military drones.[26] The argument is that when drones are deployed, no individual can be held fully morally responsible, but the military as a whole can be.

We need to distinguish a responsibility gap arising from the individual–collective issue, from the responsibility gap arising from the agency–responsibility issue of future autonomous drones. These are two different problems.[27] The individual–collective issue of collective responsibility arises in *horizontal groups*, in which members act jointly by each doing his or her part in a horizontal division of labour (cf. Shapiro, 2011, 141).[28] This contrasts with *vertical groups*, which are characterised by a hierarchical structure of power and authority. In vertical groups there are some individuals who perform

[25] See French (1979), Pettit (2001, 104–124; 2007), and List and Pettit (2011, 153–69). See Hindriks (2009) and Braham and van Hees (2011) for arguments against the plausibility and relevance of collective responsibility.

[26] Schulzke (2013, 221) writes: "Because responsibility for military decisions is so often divided, it is usually futile to attempt to blame a single actor for a given misdeed." To solve the responsibility gap problem for drones, he suggests using the existing "system of shared responsibility ... as it is already used to distribute responsible [*sic*] for the actions of autonomous human soldiers".

[27] The individual–collective issue gives rise to responsibility gaps without drones. The above case of the *Herald of Free Enterprise* is one example. Conversely, a responsibility gap arises in the drone deployment case without horizontal groups (see the argument above).

[28] The various analyses of such joint actions do not apply to hierarchical groups. They usually assume that individuals *mutually respond* to each other's behaviour and intentions, and thereby fail to represent the power structures that are characteristic of hierarchical groups (Bratman, 2014; Shapiro, 2014). See also footnote 1 above.

actions on the orders and on behalf of others. The responsibility gap of future autonomous military drones concerns only groups of this latter vertical kind.

4.1.4 Success–Failure Issue: Risky Actions

Some argue that the commander cannot be held responsible because she did not intend the drone to bomb illegitimate targets.[29] Did she not give exactly the opposite command? This consideration seems to provide alternative support for the view that the commander cannot be held responsible for the bombing. This argument appeals to the following assumption.[30]

> *Responsibility Requires Success*.
> For any a and any x, if a is responsible for x then a intended x to occur.

An action is a *failure* from the perspective of an agent if this agent did not intend the action to occur. Conversely, an action is a *success* from the perspective of an agent if this agent did intend the action to occur.

Since the commander intended for the drone to bomb only legitimate targets, the bombing is a failure from her perspective. Assuming that responsibility requires success, it follows that the commander cannot be held responsible for the bombing. On the same token, assuming that the drone intended to bomb only legitimate targets, the drone cannot be held responsible for the bombing either. Neither of them can be held responsible for the bombing. We

[29] Sparrow (2007, 70) assumes that responsibility requires success. He writes that otherwise "we simply insist that those who use [drones] should be held responsible for the deaths they cause, even where these were not intended". A variation of this principle is formulated in terms of knowledge. The idea is that the commander is not responsible for the bombing because she does not know that the drone is bombing illegitimate targets. In this vein, Matthias (2004, 175) writes: "[T]he agent can be considered responsible only if he knows the particular facts surrounding his action." A similar principle, again in terms of knowledge, is endorsed by Fischer and Ravizza (2000, 13), who write that "an agent is responsible only if he both knows the particular facts surrounding his action, and acts with the proper sort of beliefs and intentions". As the formulation by Fischer and Ravizza indicates, conditions about intentions and knowledge are closely related. Whether formulated in terms of knowledge or in terms of success relative to intentions, the principles face the same counterexamples: risky actions and negligence. We hold agents responsible despite their intentions to the contrary (risky actions) and despite their limited knowledge (negligence).

[30] Note that there is a connection between the principle that responsibility requires success and the Doctrine of Double Effect (DDE). Both principles restrict an agent's responsibility to these aspects or the effects of an action that an agent intends. Although these principles are related, the principle that responsibility requires success is much stronger than DDE. What I say against the former does not *mutatis mutandis* apply to the latter. Actions to which the DDE is relevant are both a success and a failure from the perspective of an agent. For simplicity, I focus only on actions that are either a success or a failure from the perspective of an agent.

have a further argument for a responsibility void. This is the *success–failure issue*.

There are two things to say here. First, the success–failure issue does not give rise to a responsibility gap as defined above. Second, the principle Responsibility Requires Success is false given that agents can be responsible for negligence. On the first point, the void does not arise because of drones. Hence, it does not meet the second necessary condition for a responsibility gap. Suppose that a person instead of a drone mistakenly bombs illegitimate targets. As before, the commander cannot be held responsible for the bombing. But the person cannot be held responsible for the bombing either. From his perspective the bombing was a failure too. He intended to bomb only legitimate targets. If responsibility requires success, then the person cannot be held responsible for the bombing. The responsibility void hence does not arise because of drones.

On the second point, the argument for a responsibility void due to the success–failure issue is not sound. This relates to a point that will become relevant later. Responsibility does *not* require success. The assumption above is false. Agents can be held responsible for failures. Famously, a case in point is the conviction of the Japanese general Tomoyuki Yamashita for the Manila massacre in 1946. His defence argued that Yamashita intended his troops to retreat but that he could not order them to do so because United States troops had interrupted the communication in his chain of command.[31] Effectively, the defence claimed that the massacre was a failure from the perspective of Yamashita, and that therefore he should not be held responsible for it. The court did not agree with this argument and Yamashita was convicted.

Practically all actions may turn out differently than how an agent intends them to turn out. Too often we fail in achieving what we want. This is because most actions are risky. An action is *risky* if its success is not determined by its performance. Suppose I want to remove a block from a rickety tower in the *Jenga* game. I wriggle the block out carefully but the tower collapses. My action is a failure. But it could have been a success. I believed there was a good chance that I would succeed.

We regularly hold agents responsible for their risky actions. My *Jenga* teammate may hold me responsible for the collapse of the tower. She perhaps

[31] This historic case is controversial for many reasons. Among the contested aspects is precisely whether this statement by his defence was accurate. Reel (1949) defends this line, and Whitney (1950, 5) objects to it.

may think I should have tried a different block. But let us consider a less playful example with greater moral relevance. Suppose a fighter pilot intends to bomb only legitimate targets, yet he wrongly identifies the surrendering soldiers as legitimate targets. Assuming that the pilot could have easily recognised that the soldiers had indicated their desire to surrender, then the pilot could be held responsible for the bombing even though it was a failure from his perspective. It seems that the pilot did not take proper care in making his decision to bomb what seemed to him to be legitimate targets. Most cases of negligence are of this kind. Since we regularly hold agents responsible for their risky actions, their failures, and their negligence, it is false that responsibility requires success.

4.2 Agency in Hierarchical Groups

Hierarchical groups have two characteristic features: hierarchy and control.[32] *Hierarchy* concerns a group's structure. It says that the group has different levels and that some individuals are higher up than others. The structure of the group is described by an asymmetric relation between the individuals. *Control* concerns the group's functioning. It says that those higher up in the chain of command — the superiors — control what their subordinates do by giving orders. While superiors have control over the outcomes that result from their orders, they do not bring about these outcomes themselves. Rather, any outcome is brought about by subordinates.

[32]There is a third characteristic feature of hierarchical groups that I set aside for the purposes of this chapter: *low-level overdetermination*. It consists of two assumptions. First, that for each order there is more than one individual who could carry it out. Second, that while it is open as to which individual carries out the order, it is certain that one individual will carry it out. Compare ICC (2008, §§484–518).

Characteristics of Hierarchical Groups.

Hierarchy. The group structure is described by the asymmetric relation *a has authority over b* that holds between group members. Any *a* can give orders to any *b* if and only if *a* has authority over *b*.

Control. Any *a* who gives an order has control over an outcome that is brought about by *b*, such that

1. if *a* had given orders to *b*, then the outcome would obtain, and
2. had *a* not given these orders to *b*, then the outcome would not obtain.

Hierarchical groups are not hard to find. The military comes to mind because of the wording in terms of superiors who "give orders" to subordinates. But we should not restrict the meaning of "orders" to the military sense. Instead, we include civil and friendly forms of control. Private corporations may also be hierarchical groups. Many orders do not sound like such but play a very similar role. A chief executive may exert control by announcing what she believes is the best thing to do. At other times, subordinates may try to anticipate the wishes of their superiors. Both mechanisms approximate the function of an order in different ways. There are different kinds of orders ranging from explicit, via implicit, to anticipated orders. They might then be called instructions, policies, or wishes, but each may count as an order for present purposes.

We define control in terms of two counterfactual conditionals. Counterfactual definitions are notoriously susceptible to problems of overdetermination. When in addition to one individual *a* there is another individual a^* who gives orders that bring about the *same* outcome as *a*'s order, we run into a problem. Neither *a* nor a^* would have individual control over the outcome.[33] Fortunately, we can set this problem aside. In many hierarchical groups, and certainly in their idealised examples that I consider here, the individuals have domains of control that do not overlap. That is, for any hierarchical group, if there is one individual that gives an order with an outcome, then there is no other individual that gives an order with the same outcome.[34] A hierarchical group with clear rules about who is in charge of what avoids overdetermination problems.

[33] Because each order by *a or* a^* is sufficient for the outcome to obtain, neither *a* nor a^* individually meets the second conditional of control.

[34] More precisely, we represent outcomes as sets and require that there are no intersecting outcomes without one outcome being a subset of the other.

In addition to the two characteristics of Hierarchy and Control, hierarchical groups exhibit a peculiar feature. As we have seen, in hierarchical groups we find *distant responsibility*. An agent can be held responsible for an action without performing the action herself. Furthermore "the degree of responsibility increases as we draw further away from the man who uses the fatal instrument" (Jerusalem District Court, 1961, 197). This phenomenon of *increasing responsibility* demands explanation. In addition to increasing responsibility, there are two further phenomena that any theory of agency in hierarchical groups needs to accommodate. It will be useful to attend to these phenomena first.

Three Questions.

Superiors' Agency. How are superiors agents of actions that subordinates perform?

Subordinates' Agency. How are subordinates agents of their actions when their superiors have control over the actions they perform?

Increasing Responsibility. Why does responsibility tend to increase with hierarchy?

The first question concerns the agency of superiors. When we assume that a superior can be held responsible, and we also assume that responsibility requires agency, then superiors must be agents of the actions for which they are responsible.[35] So the question is: how and in what sense are superiors agents of actions that their subordinates perform? This is the question of downward agency or of *superiors' agency*.

The second question concerns the agency of subordinates. Supposing that superiors are agents of actions that their subordinates perform, how and in what sense are subordinates agents of their own actions? This is the question of low-level autonomy or of *subordinates' agency*.[36] Let us consider each question in turn.

[35]Note that distant responsibility together with the principle that responsibility requires agency implies that the following principle (the Embodiment View of Chapter 3) is false.

Agency Requires Performance.
To be an agent of x requires doing x oneself.

This is worth noting because the principle that agency requires performance is a plausible assumption when we understand agency in a more restricted sense of "doing something" as opposed to the broader sense in use here, which is the sense of agency that is necessary for responsibility.

[36]A similar problem arises for collective responsibility and for collective agency. See List and Pettit (2011, 160–63), and Szigeti (2014) respectively.

4.2.1 Superiors' Downward Agency

There is a natural answer to the question of superior's agency that would also counter the Responsibility Void Argument: superiors control the actions of their subordinates. But this answer has been rejected by most contributions to the literature on future autonomous drones.[37] I argue that control, suitably defined, provides a theory of agency in hierarchical groups that answers the three questions.

To be the agent of an action is to control whether that action occurs. Or more specifically, I contend that to control whether an action occurs is sufficient to be an agent of that action.[38] It is important to carefully define what we mean by "control". I draw on a recent account that was developed to address a problem of mental causation to formulate the following sufficient condition for agency.

> *Agency as Control*.
> a is an agent of an action x if a has an intention such that,
> 1. if a had this intention, then x would have occurred, and
> 2. if a had not had this intention, then x would not have occurred.

Agency is a special type of control because it is *intentional control*. It is by having an intention that an agent controls whether an action occurs.[39] Both conditionals in the definition should be read subjunctively to indicate that we require more robust truth conditions than what is usually assumed. Specifically, when such a conditional has a true antecedent, we require not only that its consequent is true in the actual world, but moreover that its consequent is true in *all nearby possible worlds* in which the antecedent is true.[40]

Suppose I intend to have milk in my coffee. I open the fridge, fetch the milk, and pour it into the mug. Had I not had this intention, then these actions

[37] Consider, for example, Sparrow (2007, 71), who writes: "The use of autonomous weapons therefore involves a risk that military personnel will be held responsible for the actions of machines whose decisions they did not control." Or Matthias (2004, 177): "[N]obody has enough control over the machine's actions to be able to assume the responsibility for them."

[38] For the purposes of this chapter, we can set aside the reverse claim that control is *necessary* for agency. For similar definitions of agency see Jackson (1987, 94), Baker (2000, 149), and Alvarez (2013).

[39] An intention need not be a plan or a mental state of which an agent is consciously aware. A belief–desire pair may count as an intention for our purposes here.

[40] This is the *difference-making* account of List and Menzies (2009). See also Lewis (1973b, 27–31). As so often occurs with semantics for counterfactual conditionals, what counts as nearby must be taken as given.

would not have occurred.[41] In all nearby worlds where I intend to have milk, I open the fridge, fetch the milk, and pour it into the mug. And had I not had this intention, I would not have carried out these actions.

It should be noted that I may be an agent of what is a failure from my perspective. Just as responsibility does not require success (see above), *agency does not require success* either. Again, suppose I intend to have milk in my coffee. However, unbeknownst to me, in the fridge there is only orange juice. I mistakenly pour orange juice into the mug. I am an agent of this action because my intention to have milk in my coffee dictates that I pour orange juice in my coffee. This is a failure from my perspective because the outcome is not what I intended. Who wants orange juice in their coffee?

Control is a relation between two outcomes or events. I understand what an event is very broadly such that having an intention counts as an event, as does an omission, or the conjunction of any two events.[42] We can represent events as sets of those possible worlds in which the event occurs.[43]

With this definition of agency in hand, we can explain superiors' downward agency. Consider an example. Suppose a mafia mobster boss orders a hitman to kill a victim. The hitman kills the victim if and only if he gets the order to do so from the boss.[44] The mafia boss controls whether the death of the victim occurs. Given our understanding of agency as control, despite not performing it himself, the mafia boss is an agent of the killing. When control is sufficient for agency, then superiors are agents of the actions that their subordinates perform. In this sense, agency as control explains superiors' downward agency.

Superiors' have downward agency also when risk is involved. Suppose there is a risk that the hitman kills the wrong victim, that is, he does not kill the

[41] I use the plural here but neither here nor elsewhere in this chapter do I commit to any particular way of individuating or enumerating actions.

[42] For example, my friend's marriage and the explosion of a freight train together form a further event. For a similar principle see Jackson (1987, 93).

[43] We represent outcomes in the same way. Compare this with Lewis (1973a). Note that this is a simplification. To be precise, we would want to make events more fine-grained and to formalise them as *sets of parts of possible worlds*.

[44] We have assumed the hitman kills if and only if he gets the orders to do so. This assumption that the hitman does not or cannot resist his orders is unrealistic. But this should not obstruct our investigation. I set more realistic and more difficult cases aside for discussion later on in the thesis. For now, this unrealistic case should be taken as a model for hierarchical groups more generally and this is an idealising assumption. Furthermore, in some particularly pernicious hierarchical groups, subordinates may in fact not be able to resist orders. This account suggests that there is an action for which they can nevertheless be held responsible. The degree to which a subordinate can be held responsible will, among other things, depend on the leeway he enjoys. Measuring this precisely is beyond the aim of this chapter.

person who the mafia boss intended to be killed, but he kills a different person instead. It is then not under the control of the mafia boss whether the intended victim dies. It is not the case that the intended victim would die if the boss were to give the order. Instead, the wrong victim might die as a result of the boss' order. Nevertheless, there is an outcome that is under the control of the mafia boss. Is is under the mafia boss' control whether either one of the two victims dies. Hence, she is an agent of the outcome that either the intended victim or the wrong victim dies.

4.2.2 Subordinates' Agency and Low-level Autonomy

The question that subsequently arises is that of low-level autonomy. Intuitively, the hitman should also be held responsible. How do we explain this? There seems to be a problem (cf. Szigeti, 2014, 111). We have assumed that responsibility requires agency, we have defined agency as control, and we have supposed that the hitman does not control whether or not the victim dies. But if the hitman is to be held responsible, he must have something under his control. We answer the question of subordinates' agency by attending to a particular feature of orders.

Orders are usually underspecified or *abstract*.[45] They are made more specific by those who carry them out. There are several mutually exclusive ways that subordinates believe they can bring about an outcome that fulfils the order. Consider the couple Fred and Wilma. Fred devoutly adores Wilma, who is the six-figure earning breadwinner, while Fred is working on a doctorate. Wilma expects a nice Sunday brunch after a busy week and Fred knows this. While obviously it is better called a wish, we can interpret this as an order from Wilma to be served a brunch. There are different options open to Fred. He can make either rye bread, biscuits, or bagels. Each of these options would fulfil Wilma's wish, or so Fred believes.

> *Orders are Abstract*.
> All orders are abstract, that is, there are different mutually exclusive alternatives open to an individual b, at least one of which b believes will bring about an outcome that would fulfil the order.

The brunch may be a success or it may be a failure from Wilma's perspective. Suppose Fred makes biscuits and Wilma is delighted because she loves

[45] Different ways in which instructions are incomplete are discussed in DeMott (2014).

biscuits for brunch. Fred was right in believing that making biscuits would satisfy Wilma's wish. He could also have been wrong. He would have failed to fulfil Wilma's wish if he had made bagels because contrary to what he believed, Wilma does not like bagels for brunch.

I argue that there are two actions. While superiors control *that* a certain action occurs, subordinates control *how* they carry out the order. Since orders are abstract, they usually leave leeway in which subordinates exercise control. For these actions by which subordinates make abstract orders more specific, they can be held responsible.

Think again of the mafia mobster who orders the hitman to kill a victim. The victim must die but the hitman can choose *how* he kills the victim. He can use poison, a gun, or make it look like a traffic accident. This is something that is under his control. While he does not control that a killing will occur, he controls what kind of killing will occur.[46]

Subordinates and superiors are agents in the same sense of having something under their intentional control. A superior forms an intention, and accordingly gives an order about what outcome should occur, and she thereby controls whether this outcome occurs.[47] The mafia boss, for example, intends that a killing should occur. A subordinate forms an intention about which of the different alternative actions to pursue. By forming this intention he controls whether a poisoning, a shooting, or a traffic accident occurs.[48]

Table 4.1: Two actions in the case of the mafia killing.

Agent	**Name of action**	**Outcome**
Mafia boss	killing	that the victim is killed
Hitman	shooting	that the victim is shot

[46]Nothing of substance depends on illustrating this distinction in terms of *that* an outcome occurs and *how* it occurs. Formally, each agent intentionally controls the occurrence of the outcome.

[47]For simplicity, we assume that there is a match between intentions and outcomes and set aside failures and mistakes.

[48]There is an alternative way of answering the question of subordinates' low-level agency. We might define agency disjunctively as control *or performance*. Then we would have two agents of the *same* action. The mafia mobster boss would be an agent of the killing because she controls it. And the hitman would be an agent of the killing because he performs it. I reject this alternative for two reasons. First, it is unclear why merely performing a bodily movement is sufficient for agency in the sense relevant for moral responsibility. Second, this alternative way of accommodating subordinates' agency cannot easily answer the question of increasing responsibility.

Subordinates and superiors are agents of distinct actions (Table 4.1). A superior is the agent of an action, which has as an outcome *that an action occurs*. A subordinate is an agent of an action, which has as an outcome *how an action occurs*. In an important sense, the outcome of the action of the superior is *larger* than the outcome of the subordinate's action, which in turn is *nested* within that of the superior. Generally, in any strictly hierarchical group with n levels of hierarchy, there are n actions that are distinct but nested within each other. Let us define these notions.

> *Outcomes*.
> The outcome of an action is the set of possible worlds in which the action occurs.
>
> *Distinct Actions*.
> Two actions are distinct if their outcomes are not identical.
>
> *Nested Actions*.
> An action a is nested within an action b if and only if the outcome of a is a subset of the outcome of b.

Suppose the hitman shoots the victim. By slightly bending our use of ordinary language, we might say that the mafia boss *kills* the victim and the hitman *shoots* the victim. More precisely, the mafia boss is an agent of the action with the outcome that the victim is killed, while the hitman is an agent of the action with the outcome that the victim is shot. Since their outcomes are not identical, these actions are distinct. Furthermore, the action of the hitman is nested within the action of the mafia boss. The outcome of the hitman's action is a subset of the outcome of the boss' action. All the worlds in which the victim is shot are worlds in which the victim is killed.

In summary, with agency as control and the assumption that orders are abstract, we can see that there are distinct actions on different levels of the hierarchy nested within each other. As we go down the chain of command, subordinates exercise the leeway that is left to them by making abstract orders more specific. This is what makes them agents.

4.2.3 Increasing Responsibility

We are now in a position to answer the third question. Why does responsibility tend to increase with hierarchy? The answer is that *agency* already

increases with hierarchy. The higher up an agent is, the more things are under her control. We represent this by saying that the degree of agency that an individual wields over an outcome increases. To make this intuitive idea precise, we can define the degree of agency as a function that, intuitively, measures the size of the outcomes of actions. The size of an outcome is determined by the range of things an agent would have prevented if she were not to give an order. While much could be said to characterise this function, we need assume only that agency increases with hierarchy.[49]

> *Agency Increases with Hierarchy*.
> For any two agents of actions that are nested within each other, the degree of agency over their respective action increases from the nested action to the larger action.

With regard to a superior, this principle says that her degree of agency over her action is greater than the subordinate's agency over his respective action. In other words, if the superior were not to give her order, she could prevent a greater number of outcomes than the subordinate could by resisting to follow the order.[50] The principle says nothing about the degree of agency *between* actions that are not nested within each other.[51]

As agency increases with hierarchy, so does responsibility. This is because responsibility increases with agency.

> *Responsibility Increases with Agency*.
> The degree to which an agent can be held responsible for an action increases with the degree of her agency over an action.

[49] This principle holds only as an idealisation. As an empirical fact, control within a hierarchy is not perfect. When control is imperfect, the subordinates act outside their orders. An example would be a hitman who can resist orders. The account sketched here provides a useful model of an ideal type of a hierarchical group.

[50] The degree of agency can be said to measure the relative power to prevent outcomes. There is a different sense of agency understood not as the power to prevent but rather as the power to fine-tune the result. On this latter sense of fine-tuning, subordinates generally would have a higher degree of agency than superiors.

[51] Let me clarify Agency Increases With Hierarchy. Consider an action with a very large outcome. For example, suppose the outcome is that the Geometrisation Conjecture, a mathematical proposition, is true. The degree of agency over this action should be quite high. One might object that this is implausible because the outcome that some mathematical proposition is true seems easy to achieve. But this impression would be false. Being the agent of an action with such a large outcome is not easy at all, because it requires that the agent could have made it otherwise. You are the agent of an action that has as its outcome the truth of the Geometrisation Conjecture only if depending on what you intend, this proposition would be false.

Of course, many things matter to determine the degree of responsibility.[52] This includes for example, what an agent knows and could have known, and the reasons for which an agent acts. The function that represents the degree of responsibility must take all these considerations into account. Here, however, we are only focussing on *one* dimension of moral responsibility, namely its agential dimension.

In summary, we have a theory of agency in hierarchical groups. In contrast to *joint actions*, where there is one action with many agents, in a hierarchical group there are distinct actions on different levels of the hierarchy. And in contrast to *collective responsibility*, where no individual is fully responsible, only individuals are fully responsible.

4.3 Command Responsibility

We can see now why, when a drone is deployed by a hierarchical group like the military, no responsibility gap arises. As far as the condition of Responsibility Requires Agency is concerned, the commander can be held responsible for the bombing. The crucial premise in the Responsibility Void Argument is that the commander is *not* an agent of the bombing. With Responsibility Requires Agency it follows that the commander cannot be held responsible for the bombing. But we have seen that in the above theory of agency in hierarchical groups, this premise is false.

The commander *is* an agent of the bombing because the outcome that a bombing occurs is under her intentional control. If she were to intend that legitimate targets be bombed, then *some* targets might be bombed. And conversely, *no* targets would be bombed if she were not to have this intention and were not to give the order accordingly.

Table 4.2: The decision problem of the drone. Underlined is the actual outcome.

	Targets legitimate	**Targets illegitimate**
Bomb	Success	Failure
Don't bomb	Failure	Success

[52] Two clarifications should be made concerning the degree of responsibility. First, measuring the degree of responsibility is consistent with holding individuals *fully* morally responsible. This is because it is the degree of responsibility with respect to a specific action. Second, this measure does not allow comparing the degree of responsibility across actions. *It only measures the relative degree of responsibility between actions that are nested within each other*.

This is compatible with the view that there is something of which the drone is an agent as well. Its low-level autonomy distinguishes the drone from a mere tool. The commander's order is *abstract* because it does not specify which particular targets are legitimate and which are illegitimate. This is a decision the drone needs to make (Table 4.2). As a result, there are two distinct actions (Table 4.3). One is the action of the commander and the other is the action of the drone. While we colloquially call both actions "bombing", they are each distinct because they have different outcomes. The outcome of the commander's action is that *some* targets are bombed. The outcome of the drone's action is that *certain* targets are bombed. The drone's action is nested within the commander's larger action.

Table 4.3: Two actions in the Drone Deployment case.

Agent	**Name of action**	**Outcome**
Commander	bombing	that targets are killed
Drone	bombing	that illegitimate targets are killed

There is a final complication in the case of Drone Deployment. The outcome is a failure from the perspective of the commander. The commander intends the drone to bomb only legitimate targets, yet illegitimate targets are bombed. But is this a reason to think that the commander is *not* an agent of the bombing or that she cannot be held responsible for it? No. As I have argued, success is required neither for responsibility nor for agency. We regularly hold people responsible for negligence, ignorance, and the risky actions they undertake. Just as my action of wriggling out a particular *Jenga* block, the commander's action is *risky* because giving the order does not determine that the bombing will be successful. There is a chance that illegitimate targets will be bombed.

Sparrow (2007, 73) writes that "as machines become more autonomous a point will be reached where those who order their deployment can no longer be properly held responsible for their actions". The theory of agency in hierarchical groups shows that this is false. The notion that superiors exercise downward agency is consistent with the low-level autonomous decisions of subordinates.[53] Likewise, the fact that the drone is an agent of the bombing is not a reason for the commander to claim that she had nothing to do with it. The commander is an agent of the bombing that the drone performs.

[53] A related problem arises in the context of collective agency more generally (Pettit, 2007; List and Pettit, 2011, 160–63; Szigeti, 2014). It has an analogue in the so-called exclusion problem of mental causation (Kim, 1989; List and Menzies, 2009). In Chapter 7 I argue that the exclusion problem arises also in the context of individual agency.

In summary, the theory of agency in hierarchical groups explains two important things about the Drone Deployment case. First, it explains how the commander has downward agency over the bombing. There is an action that is under the commander's intentional control. Second, the theory thereby explains how the drone counts as autonomous. There is a further action that is under the drone's control. The drone has to decide between bombing or not bombing the particular targets it encounters. The bombing of which the drone is an agent is nested within the commander's action.

4.4 Conclusion

In this chapter I have put forward a theory of agency in hierarchical groups and exemplified it with the case of future autonomous drones. The military is only one example of a hierarchical group, however. Private corporations are another example. When Lawson (2011, 240) reflects on the question of how we can attribute moral responsibility within private corporations, he encounters the same issues with which we have grappled here.

> Our intuitions tell us the shipping clerk is less responsible than the vice-president, and it is tempting to account for this intuition by claiming that the shipping clerk's causal contributions are minimal compared to those of the vice-president. However, it is not clear how the notion of "cause" would be applied here. In what sense can we say the clerk somehow caused less than the vice-president?

My aim in this chapter has been to shed light on some of these problems. In contrast to the wording in this quote in terms of causation, I have couched my discussion in terms of agency and control; however, the lessons of the theory of agency in hierarchical groups apply generally. What motivated this investigation was the worry that future autonomous military drones would give rise to a responsibility gap. I have reconstructed an argument to back up this worry and have then developed a theory of agency in hierarchical groups. To keep things simple, I have worked with stylised examples in which subordinates sincerely try to carry out the orders they receive.[54]

[54]The real world poses all sorts of strategical problems concerning compliance and an agent's incentives. These principal-agent problems, as they are known in economics, have been set aside for present purposes.

I have identified the characteristic features of hierarchical groups and I have formulated three questions that needed to be answered. The three questions concern superiors' agency, subordinates' agency, and increasing responsibility. That is, how are superiors agents of their subordinates' actions? How do subordinates then still count as agents in their own right? And why does responsibility tend to increase with hierarchy? I have then put forward a definition of agency as control. With this definition we can answer each of the three questions. First, the theory explains in what sense superiors are agents of the actions that their subordinates perform. Superiors control what will and will not happen. Second, it explains in what sense subordinates still have autonomy and are agents of their own actions. They act in making abstract orders more specific. Finally, by introducing the concept of nested actions, this theory explains the phenomenon of increasing responsibility. Responsibility increases together with an increasing degree of control.

With this theory in hand, we can revisit the case of drones and see now where the argument went wrong. The claim that the commander is not fit to be held responsible for the bombing because she is not an agent of it turns out to be false. As far as the condition of Responsibility Requires Agency is concerned, the commander can be held responsible.[55] We have supported the legal doctrines of command responsibility and clarified how the idea that "Commanders are responsible for everything their command does" can be understood.

[55]Of course, there might be further reasons for why the commander is not responsible. I have focused only on one of several necessary conditions for responsibility. All I have argued for is that the argument as it is usually put forward, saying that the commander is not responsible because she is not an agent, is not sound.

Chapter 5

Agency as Difference-making

Responsibility and agency are closely related. Responsibility seems to require agency, that is, it requires that you have done something for which you can be held morally responsible.[1] But this view has come under attack in several ways. First, it is argued that agency is itself a normative concept and hence cannot provide non-normative grounds for responsibility. Second, it is argued that causal theories of agency are unable to explain responsibility for omissions. Third, it is argued that responsibility requires control rather than agency.[2] In this chapter I put forward an account of agency that is non-normative, that can explain responsibility for omissions despite being a causal theory and that analyses agency as a form of control.[3] I put forward the notion of *agency as difference-making*.

The starting point of this account is Aristotle's observation that the kind of agency that is necessary for responsibility consists of the ability to act together with the ability to refrain from acting. Aristotle writes, "where acting is up to us, so is not acting" (NE 1113b6).[4] An agent with this dual ability can make a difference. She can either bring about one outcome or prevent this outcome

[1] Mackie (1977, 208) calls this the "straight rule of responsibility". Williams (1994, 55) contends that one of the "basic elements of any conception of responsibility" is that someone is responsible "in virtue of what he did". Likewise, Rosen (2015, 75) writes that "to resent a person X for an act A is to think, inter alia, that X did A".

[2] For the first argument see Knobe (2003, 2006); for the second, Sartorio (2004, 2007, 2009); and for the third, Fischer and Ravizza (2000).

[3] The exact relation between control and causation is a matter of controversy. See Kment (2010); note, however, that he understands difference-making differently than I do here. Another controversial issue concerns the exact relation between control and responsibility. In particular, the topic of this paper bears on the issue of moral luck which I cannot discuss here.

[4] I rely on a recent translation by CDC Reeve (Aristotle, 2014).

by bringing about another. This idea of agency as something being "up to us" has been examined by Jackson (1987), Baker (2000, 149), and Alvarez (2013).[5]

I draw on the work of List and Menzies (2009) to formalise this Aristotelian idea in a simple propositional semantics.[6] I first introduce and develop my proposal, then illustrate it with various applications. In particular, I show that agency as difference-making identifies the mere consequences of an action, distinguishes successful actions from failures, extends to omissions and mental actions, and explains our judgments in cases involving hierarchical groups.

I present this chapter in three parts. In Section 5.1 I use the case of hierarchical groups to challenge a naïve understanding of agency and responsibility. In doing so, I locate the difference-making account among alternative proposals. In Section 5.2 I develop agency as difference-making. In Section 5.3 I illustrate it with various cases, and finally revisit the challenge from Section 5.1.

5.1 A Trilemma

Many groups have a hierarchical structure. Corporations, government agencies, and the military come to mind as examples. These groups exhibit a *vertical* division of labour (Shapiro, 2011, 141). Some members of these groups receive orders from others and carry them out on their behalf. We interact daily with such groups and each of us belongs to one hierarchical group or another. Yet, despite their prominence in our lives, they have often been overlooked in philosophical inquiries of agency.[7]

Hierarchical groups challenge our understanding of agency and responsibility. Let us revisit this case from Chapter 1.

[5] While the proposals are formally very similar, only Alvarez (2013) makes an explicit connection to Aristotle.

[6] More precisely, I put forward necessary and sufficient conditions for the *agency relation* "*a* is an agent of *x*", which holds between an individual *a* and an action *x*.

[7] There are several analyses of joint actions, which presuppose a *horizontal* division of labour and situations of "modest sociality", in which individuals mutually respond to each others' intentions (Bratman, 2014, 7). Hierarchical groups do not fit this picture (Shapiro, 2014).

> **Commanded Killing**. Suppose Anne, a military commander, commands her team to shoot an innocent civilian, Collin. The team consists of two soldiers, Bert and Ben. Bert usually shoots first. But if he were not to shoot, then Ben would. Bert goes ahead and shoots Collin.

There are two judgments about this case. The first judgment is that Anne is *not an agent* of killing Collin. This is the *agential judgment*. Intuitively, we would say that killing Collin is just not something that Anne does. The statement "Anne kills Collin" sounds false. When we take linguistic expressions as data that inform our judgments about agency, then Anne is not an agent of the killing.[8] The second judgment is that Anne *is responsible* for the killing of Collin. Responsibility here refers to the moral evaluation of an agent's past actions. Hence, this is the *moral judgment*. This judgment finds support in our intuitions, our moral practices, and in legal doctrines of command responsibility.[9]

But virtually everybody believes that agency is required for responsibility. It explains why we hold agents responsible for their own actions but not for the actions of others.[10]

> *Responsibility Requires Agency*.
> For any a and x, if a is responsible for x then a is an agent of x.

However, agential judgments, moral judgments, and the assumption that responsibility requires agency, cannot all be simultaneously true. Consider the contradiction in deductive form where x denotes the killing of Collin.

1. Anne is *not* an agent of x. (*Agential Judgment*)
2. Anne is responsible for x. (*Moral Judgment*)
3. Anne *is* an agent of x. (by *Responsibility Requires Agency*)

One of these three claims must be wrong. Which shall we give up? First, we could deny that responsibility requires agency, but keep 1 and 2. This implies

[8] Instead of relying on linguistic data, consider the following argument by elimination (for a similar argument see List and Pettit, 2011, 159–63). We assume that there is exactly one individual who is an agent of killing Collin. Intuitively, Bert is certainly an agent of the killing. After all, Bert pulled the trigger. Unless there is more than one killing, it follows that Anne is not an agent of the killing.

[9] This is not to say that Bert is not responsible. For comparisons of different legal doctrines see Danner and Martinez (2005) and Manacorda and Meloni (2011).

[10] Feinberg (1970a, 128), Fischer and Ravizza (2000, 13), and List and Pettit (2011, 158) make this assumption (or a version thereof) explicitly. The latter two call it the "control" requirement. See also footnote 1.

that Anne *is* responsible for killing Collin, despite not being an agent of this action. Second, we could deny the moral judgment that Anne is responsible for killing Collin, and keep 1 and 3. One might argue either that she is not responsible in any sense of the term or that there is some other action, such as giving the order, for which she is responsible. Third, we could deny the agential judgment that Anne is *not* an agent of killing Collin, and keep 2 and 3. This implies that Anne *is* an agent of killing Collin.

1. Deny *Responsibility Requires Agency*.
   $\rightsquigarrow$ Anne is responsible for x despite not being an agent of x.
2. Deny *Moral Judgment*.
   $\rightsquigarrow$ Anne is *not* responsible for x.
3. Deny *Agential Judgment*.
   $\rightsquigarrow$ Anne *is* an agent of x.

Each option is plausible, but comes at a cost. The first option of denying Responsibility Requires Agency reduces theoretical simplicity. A way of relaxing this condition is through disjunctive weakening. We can say that responsibility requires either agency *or* something else. Specifically, we can distinguish between two different kinds of agency.[11] The disadvantage of disjunctive conditions is their messiness. If theoretical simplicity is a virtue, then we undermine this quality by denying claim 3.

The price paid for the second option is diminished coherence with our intuitions. If we deny claim 2 and conclude that Anne is not responsible, then this is at odds with our strong intuition that Anne is responsible for the killing. To limit the damage, we can say that Anne is not responsible but she is accountable or liable for the killing. However, accountability and liability are primarily legal rather than moral concepts, and therefore it is not clear whether their analyses sufficiently reflect our moral reasoning and emotional reactions.

Our ordinary and philosophical concepts of agency come apart if we take the third option and deny 1, concluding that Anne *is* an agent of the killing. We are effectively saying that Anne is an agent of killing Collin even though we believe that the sentence "Anne kills Collin" seems false. This inconsistency with linguistic data is the price of the third option.

[11] Anne would be an agent in one sense, and Bert would be an agent in a different sense. Responsibility would require being an agent in at least one of these two senses. Fischer and Ravizza (2000) seem to propose a similar dichotomous solution. They distinguish between regulative-control and guidance-control. See also Sartorio (2007).

The trilemma shows that each escape route requires giving up an something we may value. Depending on our methodological commitments, we may lose a theoretical virtue, coherence with intuitions, or coherence with linguistic data.[12] My choice is the third option. I develop a conception of agency that judges Anne to be an agent of killing Collin, at the expense of consistency with linguistic data. I call this proposal *agency as difference-making*, and I argue that the benefits significantly outweigh the costs. In the ensuing sections, I develop the proposal at length and then discuss various applications.

5.2 Agency as Difference-making

I begin my exposition of agency as difference-making with an intuitive suggestion. As we have seen, Aristotle emphasises the idea that something is "up to us" when describing the kind of agency required for moral responsibility. When we understand agency in this way, Anne *is* an agent of killing Collin because whether or not he dies is up to her. But what exactly do we mean by saying that something is up to someone? A natural way of understanding "up to" is to understand it as "it depends on". In particular, by saying that something is up to an agent, we mean that something depends on an agent's intentions.[13] Then again, what exactly do we mean by saying that something depends on an agent's intentions?

I put forward that something being "up to" an agent should be understood as counterfactual dependence. This means that if the agent were to intend to act, then the action would occur, and if she were *not* to intend to act then the action would *not* occur.[14] Jackson (1987, 94) holds a similar view.

> Typically our actions make a difference. Had they not been performed, things would be different from the way they in fact are. ... [T]he morality of an action depends on ... the relationship between what would be the case were the act performed and what would be the case were the act not performed.

[12] This list is not complete. We might lose coherence with linguistic data not only with the third route, but also with the second route.

[13] An intention here is any mental state that plays a certain functional role. It may be a belief–desire pair, an intention-in-action, or a proximal intention (Searle, 1983, 83; Mele, 1992).

[14] Note that I state both conditionals subjunctively because they should be read as such. I discuss the reasons for reading them as subjunctive conditionals in the later more formal treatment.

This idea of difference-making or counterfactual dependence resembles a well-known necessary condition for moral responsibility: *control*. It is widely accepted that control is necessary for responsibility.[15] I propose to understand the kind of agency required for responsibility as control, and furthermore, to define control as difference-making dependence. Agency as difference-making essentially states that an action's occurrence depends on the control exerted via an agent's intentions.

> *Agency as Difference-making (informal).*
> An individual $\underline{a \text{ is an agent of } x}$ if and only if it depends on a's intentions (in the sense of difference-making) whether or not x occurs.

An individual is an agent of an action if and only if her intending something makes the difference as to whether or not the action occurs. This means that in all similar situations in which the individual has an intention, the action occurs; and if the individual were *not* to have this intention, the action would *not* occur.

A clearer perspective requires that we state the proposal more formally. In doing so, we see that agency as difference-making enables us to distinguish actions from omissions, mere consequences from the results of an action, and successful actions from unsuccessful ones. Finally, the formalisation will clarify in what sense agency as difference-making is a causal and non-normative theory.

5.2.1 Formal Setup

I formalise agency as difference-making using Lewis' (1973b) possible world semantics. In particular, I follow the exposition of difference-making put forward by List and Menzies (2009) in the context of mental causation. This context highlights that difference-making is not only a way of defining control, but also a conception of causation (Lewis, 1973a, 563).

First of all, we need to represent propositions. Following Lewis (1973b, 1986a), let Ω be the set of all possible worlds.[16] A *proposition* is represented by a subset of Ω, namely those worlds in which some fact obtains. For example,

[15] Consider, for example, Kraemer (1978), and particularly the respective "control condition" of Fischer and Ravizza (2000, 13), and also List and Pettit (2011, 158). Jackson (1987, 94) and Baker (2000, 148–57) explicitly formulate control in terms of counterfactuals.

[16] Let us set aside the question of what kind of thing these possible worlds are.

the proposition expressed by the sentence "Jane loves Mary" is represented by the set of those worlds for which the sentence is true.

Difference-making is a relation between propositions. The truth of a proposition P makes the difference to the truth of a proposition Q if and only if: If P were the case, then Q would be the case; and if P were not the case, then Q would not be the case. I use $\square\!\!\rightarrow$ as the symbol for the variable counterfactual conditional, which we read as "if it were the case that _____, then it would be the case that _____". The symbol $\neg$ denotes the negation of a proposition.[17] Difference-making is then defined as follows (List and Menzies, 2009).[18]

> *Difference-making*.
> For all propositions P and Q, P makes a difference to whether Q if and only if it is true in the actual world that
> 1. $P \mathrel{\square\!\!\rightarrow} Q$, and
> 2. $\neg P \mathrel{\square\!\!\rightarrow} \neg Q$.

These two conditions are the *positive* and *negative conditional*. It should be noted that we evaluate these conditionals in a way that differs slightly from the orthodox way of evaluating conditionals. On the orthodox way, if P is the case in the actual world and also Q is the case in the actual world, then it follows that $P \mathrel{\square\!\!\rightarrow} Q$. But here we require more. For $P \mathrel{\square\!\!\rightarrow} Q$ to be true in the actual world, we require that this connection between P and Q holds robustly. That is, in all relevantly similar worlds in which P is the case, Q must also be the case (Lewis, 1973b, 26–31). As will become clear later, it is important to evaluate conditionals in this way because our topic concerns agency. Evaluating conditionals in this way helps to formalise the intuition

[17] The technical setup in detail (Lewis, 1973b; List and Menzies, 2009). Let Ω be a non-empty set of possible worlds. Propositions are subsets of Ω. For all propositions P, call a world ω a *P-world* if and only if $\omega \in P$. Let $\neg P$ denote the negation of P defined as the set complement $\neg P := \Omega \setminus P$. For each world $\omega \in \Omega$ there exists a system of spheres around ω, which is a set of subsets of Ω, denoted $\mathbb{S}_\omega$, with the following properties. Let P, Q denote propositions and S, T denote spheres.

Nestedness: for any $S, T \in \mathbb{S}_\omega$, $S \subseteq T$ or $T \subseteq S$.
Weak Centring: for every $S \in \mathbb{S}_\omega$, $\omega \in S$.
Exhaustiveness: $\Omega \in \mathbb{S}_\omega$.
Limit Assumption: for every $P \subseteq \Omega$ with $P \neq \emptyset$, there is a set $\bigcap_{S \in \mathbb{S}_\omega : S \cap P \neq \emptyset} S \in \mathbb{S}_\omega$, which we call the *smallest P-permitting sphere* around ω.

Truth Conditions of $\square\!\!\rightarrow$: $P \mathrel{\square\!\!\rightarrow} Q$ is true in a world ω if and only if all P-worlds within the smallest P-permitting sphere are Q-worlds.

[18] This definition does not appropriately capture the meaning of "making a difference to" for every instance in which this expression is canonically used. But I do not aim to cover all canonical usages of difference-making and hence I do not need to respect all linguistic evidence.

that when an agent brings about some outcome, it often does not matter how exactly the outcome is brought about.

In order to formulate agency as difference-making we need to represent individuals, actions, and intentions. Let A be the set of individuals and X be the set of actions.[19] The agency relation *a is an agent of x* holds between elements of these two sets. The elements of these sets represent token individuals and token actions, respectively.[20] We take these sets as primitives. This way, the proposal is compatible with different theories about the ontology and individuation of their elements.

Each action is associated with a proposition, namely the event of this action occurring. We denote this proposition with $O(x)$, which we read as "x occurs" (Lewis, 1973a).[21] We do not assume that actions *are* propositions. We only claim that each action can be *represented* by a proposition that the action occurs.

Thus far we have defined difference-making, actions, and individuals. Finally we must represent intentions in a similar manner to how we represent actions. We denote the proposition that an individual a has a certain intention y with $I(a, y)$, which we read as "a intends y".[22] It should be noted that we only represent *that* an individual has a certain intention. We do not represent the content of this intention. That is, we do not assume that intentions are propositional attitudes.

5.2.2 A Formal Approach to Agency

With this formal setup in hand we can reformulate agency as difference-making as follows.[23]

[19] To allow that things other than individuals can be agents of actions, the set of individuals A will need to include further entities such as groups or sets of individuals.

[20] Action types can be extensionally defined as subsets of X (analogous to defining properties by their extension over a set of objects).

[21] Formally, this O-mapping is defined as $O : X \to \mathcal{P}(\Omega) \setminus \emptyset$. It should be noted that O is neither injective nor surjective. It is not injective because it is plausibly a many-to-one mapping, where different actions might have the same propositional extension. It is not surjective because some propositions do not represent the occurrence of an action and are hence not associated with any action.

[22] Formally, this is defined as a mapping from agents and intentional contents to propositions. Since we lack a formal representation of non-propositional intentional content and because the relevant intentions concern actions, we represent intentional content using elements of X. The mapping is defined as $I : A \times X \to \mathcal{P}(\Omega) \setminus \emptyset$.

[23] More precisely, for any $a \in A$ and $x \in X$, a is an agent of x if and only if there is a $y \in X$ such that $I(a, y)$ and $I(a, y) \boxdot\!\!\Rightarrow O(x)$, where $\boxdot\!\!\Rightarrow$ is the difference-making relation. Baker (2000, 149) puts forward a necessary condition for responsibility that is similar to agency

> *Agency as Difference-making*.
> Any a is an agent of x if and only if there is some y such that a intends y, and a having this intention makes a difference to the occurrence of x.

Since difference-making requires that both the positive and the negative conditional are true, this means the following: If a were to intend y, then x would occur (positive conditional), and if a were not to intend y, then x would not occur (negative conditional). But note that agency as difference-making does not focus on a certain intention y of the agent. Instead, it identifies the intention that makes a difference. This is because the definition says that something is an action if "there is some [intention] y such that...". In this way, agency as difference-making not only identifies the actions that are associated with an agent, it also identifies the intentions that make a difference to the occurrence of these actions.

The formal statement departs from the informal statement at two points. First, the formal version has the additional requirement "that a intends y". This condition is needed to prevent our definition from being too broad. Without this condition, any individual would be an agent of all actions that could possibly occur. However, we want to restrict agency to actions that actually occur. Consider the following example.

> **Last Cookie**. Suppose I have a jar of cookies in my office with only one cookie left. Nobody else has access to my office. If I were to intend to eat the cookie, an eating of the cookie would occur. And if I were not to intend to eat the cookie, an eating of the cookie would not occur. I forget about the jar, so I neither intend to eat the last cookie nor do I intend not to eat it.

In this situation, even though I do not actually intend to eat the cookie, my intending to eat the cookie makes a difference for the eating of the cookie to occur.[24] After all, if I *were* to intend to eat the cookie, then it would be eaten. But it is absurd to say that I am an agent of eating the cookie. We must distinguish between actions that I actually perform, and actions that I could possibly perform if I intended to do so. I am an agent of only those

as difference-making (see her principle DS). Yet she omits the positive difference-making conditional.

[24] Both conditionals are true in the actual world. The cookie would be eaten if I were to intend to eat it (positive conditional). And if I were not to intend to eat it, since nobody else has access to my office, it would not be eaten (negative conditional).

actions that I actually perform (including omissions).[25] To restrict agency as difference-making to actual actions, we need the condition "that a intends y".

Second, it should be noted that the formal definition uses two variables to distinguish between actions (x) and their intentional content (y). This reflects the possibility that an agent's intentions may differ from her actions.[26] This will be important later when we distinguish successful actions from unsuccessful ones.

5.3 Applications

In the remainder of this chapter I discuss five cases to illustrate that agency as difference-making is a useful and plausible definition of agency. First, I discuss simple actions and show how agency as difference-making distinguishes the result of an action from its mere consequences. Second, I explain how to distinguish successful actions from unsuccessful ones. Third, I discuss how agency as difference-making treats omissions. Fourth, I turn to mental actions and examine a so-called Frankfurt case as an example. Finally, I return to the Commanded Killing case from Section 5.1.

5.3.1 Results and Mere Consequences

Some actions have not just a result but also mere consequences. The result is the event to which your intentions make a difference. In contrast, mere consequences are not up to you. Consider the following case as an example (cf. Davidson, 1963).[27]

[25] Restricting agency to actual actions is motivated mainly by expositional considerations because we express the relation under consideration here "a is an agent of x". Strictly speaking, there is no substantive reason for this restriction. Since agency is only necessary but not sufficient for responsibility, it does not follow that individuals would be held responsible for their merely possible actions.

[26] This can happen in two ways. First, an agent's intentions are often more general than the resulting action. For example, suppose that any time I intend to have coffee, I get a flat white. I do not specifically intend to have a flat white. I capriciously chose a flat white when I was first confronted with the coffee menu and I stuck with my choice. My general intention to have coffee makes a difference to the action of specifically buying flat white. Another way an agent's intentions may diverge from her actions is in cases involving mistakes. Suppose I intend to switch on the light in the room. In trying to switch on the light, I stumble against the furniture and smash a vase. The light remains off. Even though I did not intend to smash the vase, I am an agent of this event.

[27] To be clear, I remain neutral on action individuation (see Appendix A). Agency as difference-making allows that a is an agent of x but not of y even though x and y both

> **Coming Home**. Suppose you intend to turn on the light. You flip the switch and turn on the light. Unbeknownst to you, a burglar is alerted to the fact that you have returned home.

It is important to note that this case is underspecified. *To settle matters of agency, we must know the proximal possibilities.* A proximal possibility is what happens in similar situations. We require information about the proximal possibilities in order to evaluate the counterfactual conditionals with which we defined agency. Applied to the Coming Home case, the proximal possibilities may be such that you are an agent of turning on the light.[28] If you were to intend to turn on the light, then the light would turn on. If you were *not* to intend to turn on the light, then the light would *not* turn on. However, the proximal possibilities may also be different. Suppose that the light may turn on even if you were *not* to intend to turn on the light. Then you would not be an agent of turning on the light because whether or not the light turned on would not be up to you.

Here is how we generally proceed in determining which actions $x \in X$ an individual a is an agent of. Let $I(a, y)$ be the event that a intends to turn on the light. So the question is: which are the actions $x \in X$ such that $I(a, y)$ makes a difference to their occurrence $O(x)$? The individual a is an agent of those $x \in X$ for which the two difference-making conditionals are true; that is, $I(a, y) \boxdot\!\!\rightarrow O(x)$ and $\neg I(a, y) \boxdot\!\!\rightarrow \neg O(x)$. Whether the two conditionals are true depends on the proximal possibilities, that is, on what would happen. Judgments about agency depend on what nearby worlds are like.

Are you an agent of alerting the burglar? It depends. Suppose that the positive conditional is true. In all nearby worlds in which you intend to turn on the light, an alerting of the burglar occurs. Whether you are an agent of alerting the burglar then depends on the negative conditional. What would happen if you were *not* to intend to turn on the light? There are three options for nearby worlds (Table 5.1). Either the burglar is alerted in *none* of them, in *all* of them, or in *some but not all* of them. If the burglar is alerted in none

occur and y is a consequence of x (see below). But agency as difference-making is not committed to the claim that x and y are distinct actions. There is still room for a metaphysics of events that focusses only on the actual world and on which x and y are descriptions of the same event. Furthermore, a more canonical statement of the case reads "a flipping of the switch and a turning on of the light occur", instead of "you flip the switch ...". I chose the present formulation because I share the worry of Bennett (1988, 6) who writes: "Some writers about events slide into a kind of philosophers' pidgin in naming events, producing such horrors as 'an event of someone's doing things'."

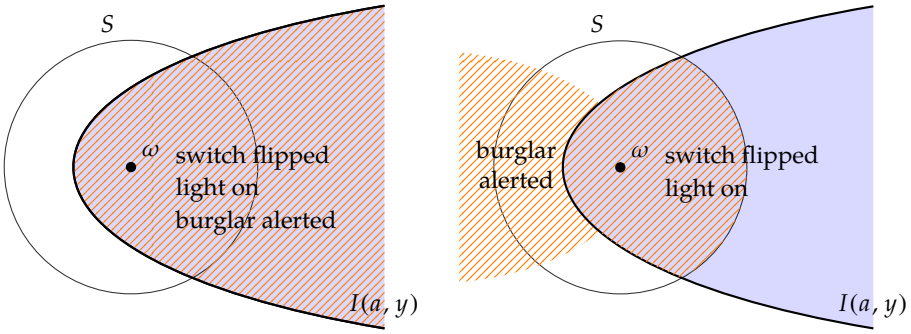
[28] There are no implicatures here concerning action individuation. I call this action "turning on the light" only for simplicity.

Table 5.1: Whether something is an action or a mere consequence depends on the proximal possibilities.

Suppose that if the agent were to intend to turn on the light, then the alerting of the burglar would occur.
If the agent were not to intend to turn on the light, then the action of alerting the burglar ...

Proximal possibility	**Alerting the burglar is ...**
would not occur	an action
would occur	not an action
might or might not occur	a mere consequence

of them, then in those nearby worlds in which you do not intend to turn on the light, the burglar is not alerted. The negative counterfactual is true ($\neg I(a, y) \boxdot\!\!\rightarrow \neg O(x)$). You are an agent of alerting the burglar (Figure 5.1a).



(a) Alerting the burglar is an action. **(b)** It is a mere consequence.

Figure 5.1: Comparison of an action and a mere consequence in proximal possibilities.

If the burglar is alerted in all nearby possible worlds, then the burglar would be alerted even if you were not to intend to turn on the light. Perhaps your partner pulls up on your driveway and the burglar is alerted before you turn on the light. In this case, you do not make a difference ($\neg I(a, y) \boxdot\!\!\rightarrow O(x)$). You are *not* an agent of alerting the burglar.

Finally, suppose that the burglar is alerted in some of the nearby worlds in which you do not intend to turn on the light, but not in others. To denote this we use the might conditional $\diamond\!\!\rightarrow$. $\neg I(a, y) \diamond\!\!\rightarrow O(x)$ and $\neg I(a, y) \diamond\!\!\rightarrow \neg O(x)$. If you were not to intend to turn on the light, then an alerting of the burglar might occur, or it might not occur. In this case, alerting the burglar is a *mere consequence* of your action of turning on the light (Figure 5.1b).

Figure 5.1 shows the proximal possibilities around the actual world ω. Whether or not a is an agent of alerting the burglar depends on whether the agent's intention $I(a, y)$ makes a difference to whether the burglar is alerted. The circle S is the sphere of nearby worlds in which the difference-making conditionals are evaluated. We define a mere consequence as follows.

> *Mere Consequence*.
> For any $x, y, z \in X$, and any a intending y such that her having this intention makes a difference to the occurrence of x:
> An occurrence of z is a mere consequence of x if and only if
> 1. if a were to intend y, then z would occur, and
> 2. if a were not to intend y, then z may or may not occur.

The mere consequence of an action always occurs when the action occurs. But when the action does not occur, the mere consequence might occur nevertheless. Since the negative counterfactual is false for the occurrence of a mere consequence, you are not an agent of the mere consequence.

This might seem problematic for agency as difference-making. Since the principle says that individuals are not agents of mere consequences, they cannot be responsible for them. Yet we can conceive of cases where agents are responsible for the mere consequences of their actions. Consider the following example.

> **Klingons**. Khan and Kor are hunting Victor to his death. While Kor is still fighting Victor's guards, Khan intends to kill Victor and succeeds. However, if Khan had not intended to kill Victor, Victor may still have died because Kor also intended to kill him.

Victor's death is in a weak sense overdetermined and is a mere consequence of Khan's actions. In overdetermination cases the negative difference-making conditional is false. It is not the case that Victor would *not* have died if Khan had *not* intended to kill Victor. Hence, Khan is not an agent of killing Victor.[29] Even though Khan is not an agent of Victor's death, intuition dictates that he is responsible for the death. This seems to be at odds with the assumption that responsibility requires agency. How can we explain this?

[29] Victor's death is *weakly* overdetermined because Kor does *not actually* succeed in killing Victor, and he succeeds only in *some but not all* nearby possible worlds. Hence, if Khan were not to intend to kill him, Victor *might and might not* die. See also my remarks on overdetermination in Section 8.1.2.

We must begin by distinguishing between two actions with slightly different results. Let x be Khan's action of killing Victor. The result $O(x)$ is that Victor is killed *by Khan*. Contrast this with the action $\hat{x}$, *someone's* action of killing Victor. The result $O(\hat{x})$ is that Victor is killed *by someone*. Khan is an agent of x but not of $\hat{x}$.[30] Hence, Khan is responsible for the action that *he* killed Victor (x), but he is not responsible for the action that *someone* killed Victor ($\hat{x}$). This distinction is easily lost under the coarseness of our natural language. The formal approach to agency allows us to speak more precisely. Each individual is an agent of an action, even in cases of overdetermination like the above.[31]

5.3.2 Success and Failure

Some actions are successful, while others are not. If I intend to turn on the light and the light comes on, then my action is successful. However, if I intend to turn on the light and the light remains off, then my action is not successful. Whether an action is successful depends on the agent's intentions. In particular, an action is successful if and only if its occurrence satisfies the agent's intentions (Searle, 1979).

The occurrence of an action x is given by $O(x)$, which is the set of possible worlds where the action occurs. Similarly, we assume that an agent's intending y is associated with a proposition given by $O(y)$, which is the set of possible worlds where the intention is satisfied. Note that I do not claim that intentions are propositional attitudes. I do not assume that the content of an intention can be represented by a proposition. Rather, I assume that whether an intention is satisfied can be represented by a proposition.

When we represent when an intention is satisfied and when an action occurs with propositions, then it is straightforward to distinguish successful actions from unsuccessful ones. Consider a variation of the Coming Home case. Suppose you know that there is a burglar outside your window and you intend to alert her to the fact that you are at home. You turn on the light and

[30] Khan does not make a difference to whether $O(\hat{x})$, whether or not Victor is killed *by someone*, but he does make a difference to whether $O(x)$, whether or not Khan is killed *by him*. $O(x)$ entails $O(\hat{x})$. That is, Khan killing Victor entails that someone kills Victor.

[31] I assume that each event P in an overdetermination case can be fine-grained to strict subsets P_i for each individual i involved in the situation, such that there is an intention of i that makes a difference to whether P_i. The event P_i is the proposition expressed by a sentence like "'P' is done by 'i'", where 'P' and 'i' are names for the event and the individual respectively. This assumption avoids circularity when the expression "done by" is given an analysis that is independent from our definition of agency. For example, we could give an analysis of "done by" in purely behavioural terms. See also Appendix A on action individuation.

the burglar is alerted. This action is *actually successful*. The actual world is as you intend, and it is that way because you intended it to be so.

However, the action might not be successful in a more general sense. Suppose, for example, there are nearby possible worlds where the burglar fails to notice that the light comes on. Hence, it is *not* the case that the burglar *would* be alerted if you were to turn on the light. She only *might* be alerted. Your action is only *possibly successful*. The result of your action is not always as you intend. In some nearby worlds where the action occurs, the result of your action is as you intend it, while in others it is not.

Ideally, an agent's action is *robustly successful*. In this case the result is as the agent intends in *all* nearby possible worlds in which the action occurs. For the opposite case, suppose the burglar is *never* alerted when you turn on the light, not even in the actual world. There is no nearby possible world in which the action occurs and the result is as you intend it. In this case, your action is *unsuccessful*.

With the formal approach to agency, we define whether or not an action is successful using set-theoretic relations as follows.

> *Success and Failure*.
> For any $x, y \in X$, and any a intending y such that her having this intention makes the difference to the occurrence of x;
> the action x is ...
>
> - actually successful if and only if $\omega \in O(y)$,
>   *i.e. the actual world is as the agent intends*;
> - robustly successful if and only if $O(x) \subseteq O(y)$,
>   *i.e. the action occurring entails that the intention is satisfied*;
> - possibly successful if and only if $O(x) \cap O(y) \neq \emptyset$,
>   *i.e. the intention is satisfied in some worlds in which the action occurs*;
> - unsuccessful if and only if x is not possibly successful.

Many actions cannot be robustly successful because they involve risk. *Risky actions* may turn out one way, but they may also turn out differently. For example, the result of flipping a coin may turn out as you intend when the coin comes up *heads* and you win. However, the result may turn out differently. The coin may come up *tails* and you lose. How a risky action actually turns out is up to nature. In risky actions, the agent's intention makes

a difference only to a disjunctive result, such as the coin landing on heads *or* tails.

There is a significant body of literature concerning the conditions under which someone does something *intentionally*. Actions that are done intentionally are a subset of all an agent's actions. Among the things you do, some things you do intentionally, and others you do unintentionally. If I mistakenly spill the glass of orange juice when reaching for the butter, this is an action of mine, albeit an unintentional one. Agency as difference-making only states conditions for agency over actions. I do not offer conditions to characterise actions that are done intentionally. However, the formal approach to agency may help to define such conditions.

5.3.3 Omissions

We can be responsible for our omissions. For example, I am responsible when I could help someone in need easily but refrain from doing so. Our intuitive grasp of omissions is firm. Yet, existing proposals on the nature of omissions are beset with problems (Sartorio, 2009; Clarke, 2012). One question, for example, is how to distinguish omissions from absences of actions. If I simply forget to buy orange juice on my way home then this is the absence of an action. In contrast, if I remember to buy orange juice but decide to go straight home instead, then this is an omission. It seems that every omission is the absence of an action, but not every absence of an action is an omission.

I put forward a way of distinguishing between omissions and absences of actions. Furthermore, I illustrate how agency as difference-making handles omissions. It turns out that it treats omissions and actions symmetrically.[32] The necessary and sufficient conditions for being an agent of x apply to actions as well as to omissions. Clarke (2010, 172) doubts whether there could be such a "unified theory". But I clarify precisely in what way both actions and omissions are "manifestations of agency" (Alvarez, 2013, 105).

The difference between omissions and absences is that the former have an intentional antecedent but the latter do not. For omissions, there is an intention that makes a difference to their occurrence. But that is not the case for absences of actions. In this sense, omissions are bound to an intention but

[32] In contrast, other theories of agency and responsibility exhibit asymmetries between "positive actions" and omissions (Fischer and Ravizza, 2000, 123–27).

absences are not. Likewise, omissions are "manifestations of agency" but absences of actions are not.

This difference between omissions and absences can be illustrated by the difference between which part of a statement about agency is negated (Table 5.2). We write "*not:*" to denote the negation explicitly.[33] Consider the statement "a is an agent of x".[34] When such a statement is about the *absence of an action*, the entire statement is negated. This is an *external negation*. If it is true that "*not:* a is an agent of x", that is, it is not the case that "a is an agent of x", then the action x is absent. That the entire statement is negated suggests not only the absence of an action but also the absence of agency. An external negation says that something is not the case.

In contrast, when a statement is about an *omission*, it is *internally negated*. If it is true that "a is an agent of *not:* x", then *not:* x is the negation of x, and *not:* x is something that a does. Hence, an internal negation suggests that there is something of which a is an agent and this something is described negatively by using "*not:* x". We may read "*not:* x" as "refraining from doing x". When a is an agent of *not:* x, we often say that a refrains from doing x.

Table 5.2: Distinction between omissions and absences of an action.

Action	Negation	Statement	Agency
Absence	external	"*not:* a is an agent of x"	No
Omission	internal	"a is an agent of *not:* x"	Yes

This canonical use of "refrain" explains why every omission is the absence of an action. We refer to omissions by describing them negatively. When I am the agent of not buying orange juice because I decide to go straight home, then we describe this omission by saying that I refrain from buying orange juice.[35]

[33] I use "*not:*" as a negation operator in English. Note that this negation is different from that denoted by the operator ¬. *Not:* operates on formulas of a language, while ¬ operates on sets of possible worlds (i.e. propositions). *Not:* returns the inverted truth value of a given formula, while ¬ returns the set-complement to Ω of a given set.

[34] It is important to consider a statement that uses the agency relation *is an agent of* explicitly. More canonical formulations of the form "she does x" pose a problem. When statements in this more canonical formulation are negated internally, as in "she does *not:* x", they are ambiguous and could refer either to the absence of an action or to an omission. This ambiguity might explain why it has been so difficult to make the distinction between absences and omissions clear.

[35] For omissions, two statements about agency are true. First, the omission statement "a is an agent of *not:* x". Second, the absence statement "*not:* a is an agent of x".

We see that there is an important difference between absences and omissions. An action may be absent without an agent doing anything. In contrast, omissions are things agents do, often described negatively using "refrain". In this sense, omissions are similar to actions. Both omissions and actions are things to which an individual may stand in the *is an agent of* relation. But what is the difference between omissions and actions?

I contend that the difference between omissions and actions is a difference in description. An $x \in X$ is an omission if and only if it can be described as the absence of another action. More precisely, x is an omission if and only if there is another $y \in X$ such that the description "a is an agent of *not:* y" is necessarily equivalent to "a is an agent of x". The two statements describe the same result. This explains why any omission x can be described as the absence of an action y, as in *not*: y, or "refraining from doing y".[36]

Agency as difference-making offers a unified conception of agency. It encompasses both actions and omissions. The formal approach to agency represents actions and omissions in the same way. Both are elements of X. The result of an action is a proposition just as the result of an omission is a proposition. Since agency as difference-making is defined in terms of propositions it treats actions and omissions symmetrically.

Table 5.3: Distinction between actions and omissions.

Actions (categorical sense)
Actions (typical sense)
Omissions

We see now that there are two senses of "action" (Table 5.3). One is the categorical sense, the other is the typical sense. First, in one sense "action" refers to something of which an individual is an agent. All elements of X, including omissions, are actions in this *categorical sense* of "action". Second, in a more specific sense, "action" refers only to *some* elements of X but not to others. In this *typical sense* of "action", actions contrast with omissions. Similarly to "omissions", "actions" in this typical sense refers to a subset of X, namely those $x \in X$ that are not omissions. This is the typical sense

[36]The result of x is $O(x)$, and the result of *not*: y is $O(\textit{not: } y)$. Just as $O(x)$ is a set of possible worlds, so is $O(\textit{not: } y)$. When x is an omission we have a y such that $O(x) = O(\textit{not: } y)$. Note that this presupposes that the given language has sufficient names for actions y.

because in this sense "action" is the name of a type of action (in the categorical sense).[37] To illustrate this, consider the following case.

> **Bystander**. Suppose Bill passes by a pond and sees a child drowning. Bill intends to stay out of the water and the child drowns. Two things are the case. First, if Bill were to intend to stay out of the water, then the child would drown. Second, if Bill were not to intend to stay out of the water, then the child would not drown.

This case is about an omission. It is up to Bill whether or not the child drowns. He is an agent of not helping the child.[38] Let x be the omission of Bill *not* helping the child and, conversely, let y be the action of Bill helping the child. Bill is an agent of x. Or in other words, he is an agent of *not*: y. Depending on our theory of action individuation, x and y may be distinct actions. But the results of x and *not*: y are identical. We can describe Bill's action in either way. The two statements "Bill is an agent of x", and "Bill is an agent of *not*: y" are different descriptions of the same result. Furthermore, both x and y are actions in the categorical sense of "action"; both are things that Bill might do. But only y is an action in the typical sense of "action".

To summarise, concerning omissions, agency as difference-making has three advantages. First, it allows us to distinguish omissions from absences. Omissions are things that individuals might be agents of. But absences are not instances of agency. The two differ in how they relate to intentions. For omissions, there is an intention that makes a difference, but absences need no such intentional antecedent. Second, agency as difference-making is a unified theory of agency. It explains the sense in which actions and omissions are both "instances of agency" (Alvarez, 2013, 107). Since both have results

[37] Since any element of X is a token action, an action type is extensionally represented as a subset of X. "Action" in the typical sense and "omission" each refer to action types.

[38] Note that this statement is internally negated. It might be objected that the Bystander case reveals a weakness of agency as difference-making. Suppose it is not the case that Bill's intention to stay out of the water makes a difference to the child drowning. Instead, it seems plausible that this intention satisfies only the positive counterfactual of difference-making but not the negative one. That is, it is not the case that if Bill were not to have this particular intention, then the child would not drown. Absent this particular intention, Bill may have some *other* intention that would keep him from saving the child and hence satisfy the negative counterfactual. It seems that agency as difference-making presupposes a particular contrastive way to identify certain relevant intentions. But this is not true. First, agency as difference-making identifies the intention that makes a difference to the occurrence of an action. Second, note that the above objection relies on an assumption about the individuation of intentions, namely, that there is some "*other* intention". Agency as difference-making leaves open how intentions are individuated. But an objection must aim at the complete and most charitable version of the theory.

and since agency as difference-making is defined in terms of these results, it treats actions and omissions symmetrically. Third, it allows us to distinguish omissions from actions. Omissions are actions in the categorical sense of "action", that is, they are things that individuals may be agents of (elements of X). But omissions are not actions in the narrower typical sense of "action". The two differ in their description. Omissions can be described as absences of actions (in the typical sense), such that the two descriptions are necessarily equivalent.

5.3.4 Mental Actions

Some actions are mental actions. For example, when you divide the amount on the restaurant bill, when you imagine seeing your partner's face, or when you quietly recite a poem, you are performing mental actions. An action is a *mental action* if its result concerns only an agent's own mental states. In contrast, *overt actions* essentially involve the movement of a body (cf. Mele, 2003, 5).[39] Agency as difference-making treats mental actions and all other actions alike, for the same reason that it treats actions and omissions alike. We have defined agency as an individual's control over a result. This definition applies to any result, regardless of whether it is the result of an overt action, an omission, or a mental action. The result of a mental action is a proposition, just as the result of any other action is a proposition.

To illustrate how agency as difference-making treats mental actions, I offer a novel interpretation of a Black and Jones case similar to those of Frankfurt (1969).

> **Black and Jones**. Suppose that Black intends for Jones to eat an apple. At first, Jones is unsure whether or not he should eat the apple. Eventually, Jones decides to eat the apple and an eating of the apple occurs. However, had Jones decided not to eat the apple, Black would have activated a secret implant that would have compelled Jones to eat the apple, and an eating of the apple would still have occurred.

Black coerces Jones covertly and conditionally. The coercion is covert because Jones is unaware of it. And it is conditional because Black coerces Jones to do as Black wants only if Jones does not do as Black wants out of Jones' own

[39] This is how overt actions are usually understood. But see Chapter 3.

volition. Who is an agent of what? The case is underspecified. We need to fill out the proximal possibilities.

As I understand the case, Black is an agent of Jones eating the apple but Jones himself is *not* an agent of eating the apple. Whether or not he eats the apple is not up to Jones. Instead, Jones is an agent of the mental action of (consciously or unconsciously) deciding to eat the apple. It is up to him whether he intends to eat the apple now. In other words, there is an interpretation of the Black and Jones case, that is, a way of filling out the proximal possibilities, such that the following four judgments are true.

1. Jones is *not* an agent of Jones eating the apple.
2. Jones is an agent of Jones deciding to eat the apple.
3. Black is an agent of Jones eating the apple.
4. Black is *not* an agent of Jones deciding to eat the apple.

Frankfurt (1969, 863) writes that "[w]hat action [Jones] performs is not up to him". If Jones were not to intend to eat the apple, Black would activate the implant and the eating of the apple would occur nevertheless. There is no intention of Jones such that having this intention makes a difference as to whether Jones eats the apple. Hence, Jones is not an agent of eating the apple. The first claim is true.

Table 5.4: Two actions in the Black and Jones case.

Action	**Agent**
Jones eating the apple	Black
Jones deciding to eat the apple	Jones

There is something else that Jones does (Table 5.4). Jones is an agent of deciding to eat the apple. We assume that this decision is a mental action.[40] We set aside whether it is made consciously or unconsciously but assume that this decision results in an intention (cf. Holton, 2009; Shepherd, 2015). Yet, because making this decision is a mental action, there must also be some earlier intention controlling that this decision occurs. Having this earlier intention makes a difference to Jones deciding whether he intends to eat the apple. If Jones were to intend to decide to eat the apple, then a decision would occur. If Jones were not to intend to decide to eat the apple, then a

[40]It seems possible to acquire intentions without forming them in a mental action. Notwithstanding that, common-sense folk psychology and much of the contemporary literature in philosophy agree that such "practical decisions are momentary intentional actions of intention formation" (Shepherd, 2015, 336).

decision of his about eating the apple would not occur.[41] Hence, Jones is an agent of deciding to eat the apple and the second claim is true.

Black is an agent of Jones eating the apple. It is up to Black whether or not Jones eats the apple. If Black were to intend that Jones eat the apple, then Jones would eat the apple, either because Jones decides to eat the apple or because Black activates the implant. Either way, the positive conditional is true. Furthermore, we suppose that if Black were not to intend that Jones eat the apple, then Jones would not eat the apple. The negative conditional is true as well. Its truth depends on the proximal possibilities, and they may be such that the negative conditional is true. Figure 5.2 illustrates the situation. Because Black's intending that Jones eats the apple makes the difference to Jones eating the apple, Black is an agent of Jones eating the apple. The third claim is true.

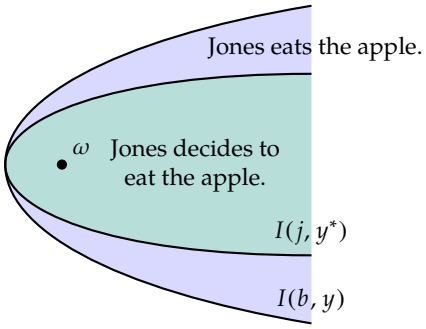


Figure 5.2: In some worlds Jones eats the apple without having decided to do so. Black intends Jones to eat the apple and Jones eats the apple in all worlds within the outer parabola $I(b, y)$. But Jones decides to eat the apple only in worlds within the inner parabola $I(j, y^*)$.

Finally, the fourth claim is true as well. Black is not an agent of Jones' decision. Whether Jones decides to eat the apple is not up to Black. The implant affects only Jones' motor control system and hence only his bodily movements. The implant does not influence what Jones intends. Hence, Black is not an agent of Jones' mental action.[42]

We have filled out the proximal possibilities. Jones intending to decide to eat the apple makes a difference to Jones intending to eat the apple. And Black intending Jones to eat the apple makes a difference to Jones eating the

[41] Just as with intentions more generally, I leave open the nature of this earlier intention that controls the formation of the later intention. This is a separate question and there are several answers available (Shepherd, 2015, 346–49).

[42] There are modifications to the Black and Jones case in which Black is an agent of Jones' decision to eat the apple. In such a modification, the implant affects not only Jones' motor control system but also his intentions and second-order intentions.

apple. Therefore, Black is an agent of Jones eating the apple, and Jones is an agent of his mental action of intending to eat the apple. In the actual world both actions occur: Jones decides to eat the apple, and Jones eats the apple. However, Jones is an agent of the former, but not the latter.

This case illustrates another crucial lesson. When the idea of agency as something that is "up to us" is taken seriously, describing what happens in the actual world to settle matters of agency is insufficient. We have the full description of a case only when we know the relevant proximal possibilities. In particular, we must know what would happen if Black were to intend something else and what would happen if Jones were to intend something else.

Let me stress that my aim is not to defend a certain interpretation of such Frankfurt cases. Instead, it was my aim to illustrate how agency as difference-making works, and to demonstrate that it is clear and useful, and that it issues plausible judgments about who is an agent of what action.

5.3.5 Hierarchical Groups

Finally, we return to the Commanded Killing case. Anne commands her team to shoot an innocent civilian, Collin. Bert, the soldier who usually shoots first, goes ahead and shoots Collin. We now see that this case combines different elements that we have encountered in the previous sections. It resembles the Black and Jones case, and involves an omission and a mere consequence.

First, note that the Commanded Killing case has the same structure as the case of Black and Jones. Figure 5.3 illustrates the situation. Bert's situation is similar to Jones' insofar as neither controls what happens. It is not up to Jones whether the apple is eaten, and it is not up to Bert whether Collin dies.[43] But unlike Jones, Bert's motor system is not controlled by a secret implant. Bert can do otherwise and resist Anne's orders.

Second, the Commanded Killing case involves an omission because Bert refrains from resisting Anne's orders. In this respect, Bert is similar to Bill in the Bystander case. Bill is an agent of the omission of not helping the

[43] Furthermore, it should be noted that these cases are similar to cases of mental causation as discussed by List and Menzies (2009). Just as there is more than one soldier who could carry out Anne's commands, in the case of mental causation, there is more than one brain state by which a mental state could be implemented.

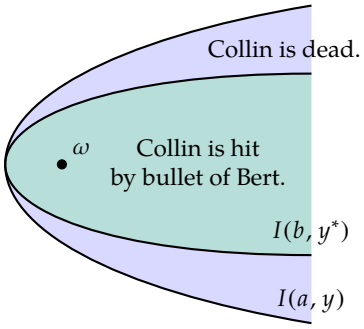


Figure 5.3: The Commanded Killing case, focussing only on Anne's intention that Collin be killed, and Bert's intention to follow Anne's orders. Anne has her intention in all worlds in the outer parabola $I(a, y)$. Bert has his intention in all worlds in the inner parabola $I(b, y^*)$.

drowning child by intending to stay out of the water, while Bert is an agent of the omission of not resisting Anne's orders by intending to follow them.

Third, the Commanded Killing case may also involve a mere consequence. While we have assumed that Ben *would* shoot if Bert were not to, more realistically, there is a chance that Ben *might not* shoot, if Bert were not to shoot. In this case, Collin dying is a mere consequence of Bert shooting him. In this respect, Bert is similar to Khan in the Klingons case. Victor's death is not up to Khan. It is only up to Khan whether or not he himself is the one who kills him. But if he does not kill Victor himself, Kor may or may not kill him. Likewise, Collin's death is not up to Bert. It is only up to Bert whether or not he himself kills Collin. But if Bert does not shoot Collin himself, Ben may or may not shoot him. Each time the result that the victim dies is the mere consequence of an action with a more specific result, namely, that the victim is killed specifically *by Bert* or *by Khan*.

This analysis of the Commanded Killing case rests on the following three claims. If we return to the trilemma discussed in section 5.1, with the first claim we take the escape route of denying the agential judgment that Anne is not the agent of killing Collin.

1. Anne is the agent of killing Collin
   (but it is not up to her who shoots him).
2. Bert is the agent of shooting Collin
   (but it is not up to him whether Collin dies).
3. Bert is the agent of not resisting Anne's orders.

It is important to keep two actions distinct. One action results in Collin's death. Specifically, the result is that Collin is killed *by someone*. The agent

of this action is Anne. We call this action the "killing". The second action has the result that Collin is killed *by Bert*. Bert is the agent of this second action, which we call the "shooting". Bert's action is an omission. We can describe his action as the absence of another action, that is we can say that Bert does not resist Anne's orders. By performing one action, he at the same time refrains from doing another.

The arguments to show that these three claims are true parallel the arguments in the respective sections above. Each time, the truth of a claim depends on the proximal possibilities, that is, on what would happen.

With the first claim: Anne is the agent of killing Collin. There is an intention of Anne such that her having this intention makes a difference to the occurrence of the result that Collin is killed *by someone*. If Anne were to intend for Collin to be shot by someone, then Collin would die. If she were not to have this intention, then Collin would not die. However, Anne has no control over which soldier shoots. Anne is the agent of killing Collin. But she is not the agent of an action with a more specific result, such as Collin being killed *by Bert*.

With the second claim: Bert is the agent of shooting Collin, that is, of the action with the result that Collin is shot *by Bert*. There is an intention of Bert, for example, that Bert intends to follow Anne's orders, which makes a difference to the occurrence of this result. The more general result that Bert is killed *by someone* is a mere consequence of Bert shooting Collin. It would occur if Bert were to shoot, but it may or may not occur if he were not to shoot. If Ben were to shoot, it would occur. But if all soldiers were to collectively resist Anne's order, or if Anne were not to command Collin to be shot, it would not occur.

With the third claim: Bert is the agent of an omission. When he shoots, he refrains from resisting Anne's orders. His shooting, while being an action in the typical sense of "action", is also an omission. In all worlds in which Bert shoots, he is not resisting Anne's orders. The description "Bert is the agent of shooting Collin" is necessarily equivalent to "Bert is the agent of not resisting Anne's orders to shoot Collin". The second statement is internally negated. Hence, Bert's typical action of shooting Collin is an omission. His individual resistance could have been the first step towards a collective mutiny. Bert is the agent of not contributing to collectively overturning Anne's order.

In the light of this, let us reflect again on the challenge posed in Section 5.1. Initially, it seemed that we needed to jettison the cherished principle that Responsibility Requires Agency in order to uphold our intuitive judgments about agency and responsibility. Against this backdrop, agency as difference-making yields a welcome result. Both Anne and Bert may be responsible for their actions as far as Responsibility Requires Agency is concerned. Agency as difference-making identifies an action for Anne as well as for Bert. It is up to Anne that Collin is killed by someone. It is up to Bert whether Collin is killed by him. Our judgments of moral responsibility should not miss out on the clarity that this formal approach to agency has to offer.

5.4 Conclusion

In this chapter I have put forward agency as difference-making. It consists of necessary and sufficient conditions for the agency relation "a is an agent of x", which holds between an individual a and an action x. The proposal is that an a is an agent of an action x if and only if there is an intention of a that makes a difference to the occurrence of x. Aristotle's suggestion that in acting, something is "up to us", inspired the notion of agency as difference-making. I have formalised this suggestion in a modified possible world semantics (Lewis, 1973b; List and Menzies, 2009).

I have illustrated the merits of this proposal in discussing five kinds of cases. First, I have distinguished the results of actions from their mere consequences. Second, I have defined different senses in which an action can be successful. Third, I have investigated how agency as difference-making elucidates the phenomenon of omissions. Fourth, I have illustrated that agency as difference-making extends to mental actions. Finally, I have shown how agency as difference-making makes sense of agency on different levels, as we find it in hierarchical groups.

Agency as difference-making is non-normative. It can distinguish the mere consequences from the results of an action by only using information about proximal possibilities and without having to appeal to normative considerations. We have also seen that agency as difference-making can explain responsibility for omissions despite being a causal theory in the tradition of Lewis (1973a). Finally, by using the notion of difference-making, it analyses agency as a form of control.

Most of the work still remains to be done. The semantics of Lewis (1973b), in which I have formulated agency as difference-making, has various limitations. It would be beneficial to have a framework with more structure to make sense of the idea that actions occur at a particular location and at a particular time. Another issue related to this choice of formal framework is the question as to what extent agency as difference-making faces counterexamples in overdetermination cases. Furthermore, the proposal so far sidesteps questions about action individuation and the ontology of actions, which are central topics in philosophy of action. Finally, I have not discussed the problem of so-called deviant causal chains. Although I believe that agency as difference-making has the resources to solve this problem, I cannot sketch the solution here. Rather, I need to leave open how agency as difference-making relates to these issues, but these relations must be worked out.

Chapter 6

Omissions

In this chapter I put forward a new solution to the *problem of profligate omissions*. The problem is that some definitions of causation identify any omission that could have prevented an effect as a cause, which leads to counterintuitive results. The solution is to strengthen the counterfactual dependence condition of causation and to weaken the centring condition of the semantics. In contrast to existing solutions, this new solution does not appeal to normative, epistemic, pragmatic, or metaphysical considerations.

This chapter extends a distinction that I have drawn in Chapter 5. In Chapter 5 I distinguished omissions from absences. Omissions are things to which intentions make a difference but absences are merely events that do not actually occur. I now apply the idea behind this distinction to a problem in the metaphysics of causation. But I must warn the reader. In this chapter, I draw the distinction in different terms. I do not introduce it with the help of agency as difference-making and I do not talk of "absences". Instead, I will discuss a condition named "Strong Dependence". This is the difference-making relation under a different name. As before, omissions are things that stand in this relation. Apart from this terminological discontinuity, this paper provides an independent motivation for difference-making, that is, for Strong Dependence. I introduce a new problem, which arises in discussions about causation, to suggest that difference-making helps to solve it.

6.1 From Weak to Strong Dependence

Common sense tells us that omissions can be causes. Suppose I promise to water your plant, but I do not water it, and the plant dries up and dies (Menzies, 2004; McGrath, 2005). My omission of not watering the plant causes its death. The following principle agrees with this common-sense judgment.[1]

> *Weak Dependence*.
> c is a cause of e if
> 1. had c not occurred, then e would not have occurred.

My omission c not to water the plant is a cause of the plant's death e because had I watered the plant, then the plant would not have died.

It turns out that Weak Dependence is too weak because it identifies too many causes. The Queen of England did not water your plant either. Had she watered the plant, then the plant would not have died. Hence, her omission is likewise a cause of the plant's death. According to this argument, the omission of anything that could have prevented the plant's death is a cause. This is the problem of profligate omissions. It is a widely recognised problem for Weak Dependence as a sufficient condition for causation (Lewis, 2000, 196; Menzies, 2004; contributions in Collins et al., 2004a; McGrath, 2005).

I put forward a new solution to this problem. I propose to strengthen the dependence condition and to weaken a condition about how counterfactual conditionals are evaluated. This proposal builds on the innovative approach by List and Menzies (2009) to address the exclusion problem for mental causation. I extend their proposal to causation more generally. When applied to the problem of profligate omissions, the result is that my omission of not watering your plant is a cause of the plant's death but the Queen's omission is *not*.[2] I discuss the details of the solution later. First, consider the following strengthened dependence condition.[3]

[1] Throughout the chapter I assume that the respective c and e actually occur.

[2] The solution here is similar in spirit to what McGrath (2005, 134–36) discusses as "proximate would-be prevention", but it is importantly different. First, the causal condition is substantively strengthened. Second, the assumptions of the semantics are different (McGrath talks of the "closest world to the actual world" and assumes uniqueness and strong centring). Finally, the solution here is simpler because it operationalises the notion of a "proximate would-be preventer", instead of taking it as a primitive. Whether McGrath's counterexamples against proximate would-be prevention can be turned against my solution *mutatis mutandis* merits a separate investigation.

[3] Strong Dependence is equivalent to Difference-making as defined in Chapter 5. I use the name "Strong Dependence" in this chapter to make the contrast with the alternative Weak Dependence condition clear.

Strong Dependence.
c is a cause of *e* if

1. had *c* not occurred, then *e* would not have occurred, and
2. had *c* occurred, then *e* would have occurred.

The first condition of Strong Dependence is identical to Weak Dependence. Hence, the first condition is true as before. Had I watered the plant, it would not have died. The second condition is true as well. Had I not watered the plant, then it would have died. Hence, Strong Dependence agrees with common sense in ruling that my omission is a cause of the plant's death.

Next I turn to the role of the Queen. According to Strong Dependence, the Queen's omission is not a cause of the plant's death. The first condition holds because the plant would not have died if the Queen had watered it. The second condition, however, is false. If the Queen had not watered the plant, then I perhaps might have shown up as promised and watered the plant.[4] There are nearby possible worlds where the Queen does not water the plant but I do. In these worlds the plant would live despite the Queen's omission of not watering it. In contrast to Weak Dependence, Strong Dependence agrees with common sense on both counts and rules that the Queen's omission is *not* a cause of the plant's death. By relying on Strong Dependence, omissions are not as profligate as they were before.

This solution is superior to existing proposals because it is simpler than any other alternative. It does not rely on normative considerations about what is normal, as suggested by McGrath (2005). This solution furthermore does not rely on epistemic considerations to distinguish between more or less salient omissions, as does the solution put forward by Bennett (1988, 133). Moreover, it does not rely on pragmatic considerations such as what causal claims are appropriate to assert in a context, as Lewis (2000, 196) proposes; nor does it rely on how the truth of causal claims depends on contextual parameters, as Menzies (2004) suggests. Finally, unlike Dowe (2010), this solution does not need to impose the metaphysical condition that cause and effect need to be proportional, which is based on the distinction between determinates and determinables. Instead, this solution simply strengthens Weak Dependence by adding a second counterfactual conditional.

[4]With the Weak Centring condition this so-called inner possibility (together with the assumption that if I were to water the plant then it would not die) entails the falsity of the counterfactual that if the Queen were not to not water the plant, it would die.

6.2 Representing Proximal Possibility

I will now go over this solution in more detail to introduce the second part to the solution. This part consists of weakening an assumption about how we evaluate the counterfactual conditionals by which we make judgments about causation. In order for Strong Dependence to deliver the results in line with common sense, we need to change the semantics of counterfactuals. Consider the following four statements.

(1) If I were to water the plant, it would live.
(2) If I were not to water the plant, it would die.

If statements 1 and 2 are true, then according to Strong Dependence my omission of not watering the plant is a cause of the plant's death. Similarly, the Queen's omission of not watering the plant is a cause of the plant's death if the following two statements are true.

(3) If the Queen were to water the plant, it would live.
(4) If the Queen were not to water the plant, it would die.

Previously I have argued that statement 4 is false, because the plant might live if the Queen did not water it. While this line of reasoning is intuitively plausible, it goes against the usual way of evaluating counterfactuals (Lewis, 1973b, 27–31). According to the standard semantics of counterfactual conditionals, statement 4 is true.[5] To solve the problem of profligate omissions, we need to tweak the semantics for counterfactual conditionals. I advance a slightly modified version of the standard semantics to address the problem of profligate omissions. Under this amended semantics, statement 4 is false.

I take it for granted that it is *more likely* that I water your plant than that the Queen waters your plant. We can say that the possibility that I water your plant is a *proximal possibility*. In contrast, the possibility that the Queen waters your plant is only a *remote possibility*. This idea can be made precise in possible-worlds terminology as follows. Let there be a set of worlds very similar to our world and call these the nearby worlds or the proximally possible worlds. In *some* of these worlds, I water your plant. But the Queen waters your plant in *none* of these nearby worlds. There surely are possible worlds in which the Queen waters your plant but they are not nearby. All

[5] Contrary to what is usually assumed, I deny that from A & B it follows that $A \mathrel{\Box\!\!\rightarrow} B$ ($\Box\!\!\rightarrow$ denotes the would-counterfactual conditional). If this were a valid inference, statement 4 would follow from the fact that the Queen does not water the plant and the plant dies.

possible worlds in which the Queen waters your plant are outside the set of proximally possible worlds.

The nearby worlds are very similar to our world. There are differences between our world and each proximally possible world but these differences might be negligible. They might be small enough to put worlds in which I *do not* water the plant (such as the actual world) in the same set as possible worlds in which I *do* water the plant. Or as David Lewis (1973b, 29) puts it:

> Perhaps our discriminations of similarity are rather coarse, and some worlds different from i are enough like i so that such small differences as there are fail to register.

I submit that when it comes to causation, our discriminations of similarity are rather coarse.[6] In sum, while I assume that, in the set of proximal possibility, there are possible worlds in which I *do not* water the plant and there are possible worlds in which I *do* water the plant, crucially, there are no proximally possible worlds in which the Queen waters your plant.[7] Such royal botanic interventions take place only in remotely possible worlds.

Assumption I. It is proximally possible that I water the plant.
Assumption II. It is not proximally possible that the Queen waters the plant.

In making these stipulations we commit ourselves to the condition of Weak Centring and deny what is called Strong Centring. Weak Centring means that the closest sphere of possible worlds around the actual world — let us call this the *sphere of proximal possibility* — may include more worlds than just the actual world. Formally put, let Ω be a non-empty set of possible worlds and let $\omega \in \Omega$ be the actual world. Let $\mathbb{S}_\omega$ be a *system of spheres of similarity* relative to the actual world ω, that is, a set of subsets of Ω. It is *weakly centred* if and only if the actual world ω is a member of each of the spheres in the system.[8] In addition, I assume the system satisfies the usual

[6] I do not claim that Lewis would agree with this idea.

[7] By proximal possibility I mean what Lewis (1973b, 30) calls "inner modalities". Whether assumptions I and II are plausible depends crucially on how coarse-grained we take the similarity relation over worlds to be. When are two worlds similar enough to be included in the same sphere? I assume that the similarity is coarse-grained enough to make assumption I true but not so coarse-grained to make assumption II false.

[8] Formally defined as follows. **Weak Centring**: for every $S \in \mathbb{S}_\omega$, $\omega \in S$.

conditions of Nestedness, Exhaustiveness, and the Limit Assumption.[9] I furthermore assume that statement 1 is true.

Assumption III. (1) If I were to water the plant, it would live.

The truth conditions for counterfactuals such as in statements 1 to 4 are as follows. I use the symbol $\Box\!\!\rightarrow$ to indicate a would-counterfactual. Its official reading is "if it were the case that _____, then it would be the case that _____" (Lewis, 1973b). A and B are propositions (subsets of Ω).[10]

> *Truth Conditions for Counterfactuals*.
> $A \Box\!\!\rightarrow B$ is true at world ω if and only if
> 1. There is no A-world in any $S \in \mathbb{S}_\omega$, or
> 2. In the smallest A-permitting sphere in $\mathbb{S}_\omega$, every A-world is a B-world.

Using assumptions I to III in the standard semantics for counterfactuals with weak centring, we can show that statement 4 is false.

Proposition. It follows from assumptions I to III that statement 4 is false.

The first condition of the truth conditions is false for statement 4. There are worlds in which the Queen does not water the plant (by Assumption II) and they are included in the system of spheres (by Exhaustiveness). The second condition of the truth conditions is false as well. There are nearby worlds where the Queen does not water the plant and yet the plant does *not* die. The smallest sphere that permits the antecedent of the counterfactual in statement 4 to be true is the sphere of proximal possibility. This is because all worlds in the sphere of proximal possibility are worlds in which the Queen does not water your plant (by Assumption II). But there are worlds in this sphere where the plant lives. This is because it is a proximal possibility that I water your plant (by Assumption I). And if I do, then the plant will live (by Assumption III). Hence, there are worlds in the sphere of proximal possibility in which the Queen does not water the plant but the plant lives. Statement 4 is false. Figure 6.1 illustrates the situation.

[9] Defined respectively as follows.

Nestedness: for any $S, T \in \mathbb{S}_\omega$, $S \subseteq T$ or $T \subseteq S$.
Exhaustiveness: $\Omega \in \mathbb{S}_\omega$.
Limit Assumption: for every $A \subseteq \Omega$ with $A \neq \emptyset$, there is a set $\bigcap_{S \in \mathbb{S}_\omega : S \cap A \neq \emptyset} S \in \mathbb{S}_\omega$.

[10] An A-world is a world in which A is true. The phrase "smallest A-permitting sphere" in the second condition means the sphere of worlds closest to the actual world that contains a world that has at least one A-world as its member. The limit assumption assumes that there is such a sphere for every A.

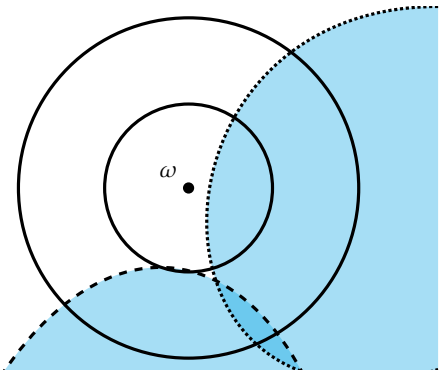


Figure 6.1: Nearby possible worlds around the actual world ω.

In Figure 6.1, the dot shows the actual world ω in the space of possible worlds. The inner circle is the sphere of proximal possibility. The shaded region covers worlds in which the plant lives, while the white region covers worlds in which the plant dies. To the right of the finely dotted line are worlds in which I water your plant. Below the coarsely dashed line are worlds in which the Queen waters your plant.

It should be noted that in all worlds in which I water the plant, the plant lives (statement 1 is true). Moreover, in all worlds in which the Queen waters the plant, the plant lives (statement 3 is true). In all worlds in the inner circle in which I do *not* water the plant, the plant dies (statement 2 is true). However, there are worlds in which the Queen does not water the plant but the plant still lives (statement 4 is false). These are worlds in the shaded area in the inner circle to the right of and above the coarsely dotted line. In these worlds, I water the plant and the plant lives.

By strengthening dependence and weakening centring we can make logical sense of the common-sense judgment that my omission of not watering your plant is a cause of the plant's death, but the Queen's omission is not.

6.3 Harder Cases

I argue that the solution extends to harder cases. One might object that the case of the Queen not watering your plant is too easy because the Queen is unlikely to water your plant. It is merely a very distant possibility. But how about the possibility that my next door neighbour waters the plant? This certainly seems to be a proximate possibility, because she is as distant from your plant as I am. So if my omission of not watering your plant is a cause of

its death, my neighbour's omission of not watering it either should be a cause too. It seems like the problem of profligate omissions returns for proximally possible omissions.

But this objection conflates modal distance (as described by the similarity relation in our model) with spatial distance. Although it might be true that I am as far away from the plant as my neighbour, it is nevertheless more likely that I water your plant than that she does — especially since I promised to do so. As such, this harder case is one that the Strong Dependence solution should also be able to handle.

Yet, there might be even harder cases. For example, suppose I promise to water your plant but, unlike my neighbour who has an innate aptitude for horticulture, I have not the slightest inclination to water it. I believe the Strong Dependence solution is still capable of delivering adequate results. But to go into more detail is to miss the important point: even if the Strong Dependence solution were to give counterintuitive results in some harder cases, it would nevertheless solve the problem of profligate omissions for almost all omissions. As long as it is proximately possible that someone waters the plant, my solution puts up a boundary that keeps omissions in check.[11] Even if harder cases were to succeed, then all remaining profligate omissions would be only within the sphere of proximate possibility, but not anymore outside of it. The sphere of proximate possibility limits omissions' profligacy.

What counts as a proximal versus remote possibility? This is a question that I have left open. By amending the semantics of counterfactual conditionals, I have made room in our semantics to capture the intuitive distinction between proximal and remote possibilities. But I have not given an account of how the distinction should be drawn. In this sense, the Strong Dependence solution is merely a formal framework that can be interpreted and employed in different ways. It is even flexible enough to draw the line between the proximate and remote possibilities on a case-by-case basis.

Compared to some of the existing solutions, Strong Dependence is a generalisation. Depending on how we interpret the formal framework — in

[11] The crucial assumption of this solution is that there are some proximally possible worlds in which the plant does not die. Otherwise, the problem of profligate omissions would return. But we can nevertheless relax Assumption I. That is, the solution does not require that it is proximally possible that *I* water the plant. As long as the crucial assumption is satisfied and the plant's death is not proximally necessary, the Strong Dependence solution can deal with several variations of the original case.

particular the similarity relation — we may recover some of the existing solutions. For example, we could interpret the similarity relation in a way that makes proximal possibilities represent what is normal. More specifically, we could understand "normal" either in an epistemic sense as what is objectively likely, or we could understand it in a normative sense as what should be the case. Under these interpretations we would get solutions similar in spirit to that of Bennett (1988, 133) and McGrath (2005), respectively.

The formal framework could also be employed as a solution to the problem of profligate omissions on the level of pragmatics. What counts as proximate possibilities may then depend on the context in which statements about causation are made.[12] The similarity relation could be interpreted as encoding either what statements are appropriate to assert, or what statements are true in a context. In turn, we would get solutions similar in spirit to that of Lewis (2000, 196) and Menzies (2004), respectively.

With this flexibility, the proposal here can be seen as generalising some of the existing rival solutions. If we understand proximal possibility as representing what is normal, then it is true — as some have suggested — that judgments about causation depend on what is normal. Furthermore, we can understand "normal" either in an epistemic sense as what is objectively likely, or we can understand it in a normative sense as what should be the case. In fact, there are many different so-called flavours of modality. The proposal here can represent any of these flavours with a corresponding similarity relation over the set of possible worlds.

6.4 Conclusion

In conclusion, we have a new solution for the problem of profligate omissions. It consists of two parts. First, we strengthen the sufficient condition for causation from Weak Dependence to Strong Dependence. To do so we add a second counterfactual. Second, we modify the standard semantics in which these counterfactuals are evaluated. They are evaluated in a system of spheres of possible worlds that is weakly centred on the actual world. I have presented a plausible situation in which my omission is a cause of the plant's death but the Queen's omission is not. In contrast to existing solutions, this novel solution does not appeal to normative, epistemic, pragmatic, or metaphysical considerations.

[12] Consider here the semantics of Kratzer (1977).

Chapter 7

Activity and Exclusion

The aim of this chapter is to reconcile two seemingly opposing accounts of agency known as the agent-causal and event-causal accounts. These two accounts have much in common. Both formulate necessary and sufficient conditions for something to be an action, and both accounts put these conditions in terms of causation. The two accounts differ, however, in *what* they identify as the cause of an action's results. While the agent-causal account says that the results of actions are caused by the agent herself, the event-causal account, in contrast, states that the results of actions are caused by the agent's mental events. This latter account faces the *problem of agential activity*. In focussing on the agent's mental events, an event-causal account neglects the fact that the agent herself is actively doing something, or so argue proponents of the agent-causal account. I suggest that both accounts of agency may be getting it right. They describe causation on different levels and there need not be any rivalry between them.

I argue that the alleged conflict between agent-causal and event-causal accounts of agency resembles another problem. An analogous conflict arises for some theories of the mind over identifying the causes of behaviour. Is behaviour caused by mental states or by neurophysiological states? Given how they are related, it cannot be both. This is known as the *exclusion problem*. Essentially, we have causation on different levels. On one level we have neurophysiological states, and above that is a level with intentions and beliefs. The question is whether each respective causal claim is true, or whether causation on one level excludes causation on another level. I argue that the problem of agential activity is likewise about causation on different levels. I formulate

the problem of agential activity as an exclusion problem. Recognising it as such helps to resolve the problem of agential activity.

This chapter comes in three parts. In Section 7.1 I describe the problem of agential activity. In Section 7.2 I describe the exclusion problem. In Section 7.3 I reformulate the problem of agential activity as an exclusion problem.

7.1 The Problem of Agential Activity

Suppose you want to reach for a cup that sits in front of you on the table. You reach for it and grab it with your hand and the cup is in your hand as a result. We take it for granted that this result has a cause. There are at least three candidate causes. First, there is the individual. You, the agent, cause the cup to be in your hand. Supported by linguistic data, this is the claim to which proponents of the agent-causal account of agency subscribe.[1] Second, there are the agent's mental events. Your intentions cause the cup to be in your hand. Here, the mental event of having an intention counts as a cause. This second claim is put forward by proponents of the event-causal account of agency. Third, there are the agent's neural events or brain states. Neural events in your motor cortex and in your neurophysiological system cause the cup to be in your hand. Neuroscience explains behaviour with such a causal claim.

Proponents of each candidate causal claim might consider one of these candidate causes to be the only appropriate cause. I suggest we do not need to take sides or rule out any of these three causal claims as false. Instead, we should try to make sense of them and evaluate whether or not they conflict. To do so, I put forward a framework to represent these claims formally.

I suggest that each respective cause operates on a different level. In particular, I propose that we consider an agent as a system with three levels. The *individual level* rests at the top, followed by the *intentional level*, where we find individual's mental events, and the *implementation level* below, where our neurophysiological system operates, involving patterns of neurons firing in

[1] We describe what happens by saying "you take the cup". In such descriptions you count as the agent of taking the cup in the linguistic sense (a semantic concept). This may be taken to support the claim that you are an agent of taking the cup in the philosophical sense. Compare Chisholm (1976a, 199): "[S]tatements [such] as 'Jones killed his uncle' and 'Jones raised his arm' are sometimes true; ... they imply that there is a certain event or state of affairs that an agent caused or brought about." See also Steward (2012, 200).

the brain. Consider now the following three statements. Each identifies a cause.

1. **Individual level**: You (the agent) cause the cup to be in your hand.
2. **Intentional level**: Your intention to reach for the cup causes the cup to be in your hand.
3. **Implementation level**: Neural events cause the cup to be in your hand.

According to the first claim, the agent herself causes the result of the action. According to the second claim, there is a mental event that causes the result of the action.[2] We remain open about what "intention" means exactly and just take it for granted that it refers to some appropriate entity for which different theories have different names. It may be called a desire, a belief–desire pair, an intention-in-action, or a proximal intention (Searle, 1983; Mele, 1992). Finally, the third claim says that the result of an action is caused by neurophysiological events in your brain and nervous system. Distinguishing these levels helps us to distinguish different accounts of agency.

Different accounts of agency formulate different conditions for something to be an action. We distinguish them by the level on which they locate the cause of an action's results. An *agent-causal* account locates the cause of an action's result on the individual level. It says that something is an action if and only if its results are caused by you, the agent. In contrast, an *event-causal* account of agency locates the cause of an action's results on the intentional level. It says that something is an action if and only if its results are caused by a mental state, such as a desire or an intention, in the right way.[3] Since the event-causal account is more popular than the agent-causal account, it is sometimes called the standard story of action. This standard story faces the problem of agential activity, or so argue its opponents. Hornsby (2004, 2), for example, puts it as follows.[4]

[2] I remain neutral about ontological categories and use them interchangeably for now. I write "mental *state*" or "mental *event*" and "*properties* of your neurophysiological system". This might suggest that the claims refer to specific ontological categories (states or properties as opposed to, for example, substances or events). They are not meant to do so here.

[3] Note that I do not claim that these are conditions for *intentional* actions or for doing something *intentionally*.

[4] Recently, Alvarez and Hyman (1998, 227–28), O'Connor (2005), Mayr (2011, 6–13), Steward (2012, 197–247), Buckareff (2012), and Pereboom (2014a,b) raise the agential activity problem. Strictly speaking, Pereboom (2014b, 32) raises a different but similar concern. He discusses the activity problem as a challenge for accounts of responsibility (as opposed to agency) in the context of free will.

> *[T]he story leaves agents out*. Human beings are ineliminable from any account of their agency, and, in any of its versions, the standard story is not a story of agency at all.

The concern is that the agent herself plays no role in the event-causal account, because it claims that results of actions are caused by mental events. Event-causal theories therefore fail to satisfy the crucial desideratum, or so claims Hornsby, to account for an agent's activity. She continues: "[A]gency cannot be portrayed in a picture containing only psychological states and occurrences and no agent making any difference to anything" (Hornsby, 2004, 12).

Several authors put forward this agential activity problem. Aguilar and Buckareff (2010, 12–14) provide an overview. They call it the problem of the passive or absent agent.[5] They say that in an event-causal account either the "agent [is] a passive participant in the production and control of an action" or it is absent from the account altogether. Each time the concern is that an event-causal account of agency fails to adequately account for an agent's activity. Event-causal accounts dislocate the relevant cause of an action's results.

Melden (1961), Taylor (1966), and Goldman (1976) are early proponents of the agential activity problem.[6]

> It is futile to attempt to explain conduct through the causal efficacy of desire — all that can explain is further happenings, not actions performed by agents. The agent confronting the causal nexus in which such happenings occur is a helpless victim of all that occurs in and to him (Melden, 1961, 128).

> [W]hen prior events alone cause us to move in one way or another, we are passive with respect to the resulting motions. We are, on such occasions, moved rather than self-moving. We are truly active only when we are ourselves the ultimate causes of what

[5] In fact, Aguilar and Buckareff (2010) distinguish two versions of the agential activity problem. I am focussing on a part of the critique that the two versions have in common.

[6] Some authors count Chisholm (1976b, 86) as another source of the agential activity problem. However, there are two problems with this attribution. First, it is not clear to what extent Chisholm subscribes to the objection (he presents it in quotation marks). Second, he states the objection so generally that it could well be interpreted as an instance of the old compatibilist concern about how responsibility is possible in a deterministic world, which is a different problem. While the agential activity problem is a concern about *levels*, the compatibilist concern is about *possibilities* (is there, and if so in what sense; does there need to be, and is there a way things could have been otherwise?).

> we do. Our causing our actions, then, must be something that is ontologically fundamental (Taylor, 1966).
>
> [T]here is the problem of agent-causation versus want-and-belief causation. If the acts of an agent are caused by his wants and beliefs, how can *he*, the agent, be considered their cause? (Goldman, 1976, 81).

Note how Goldman (1976) suggests that the choice between agent-causal and event-causal accounts of causation is exclusive ("versus"). Most of the literature accepts this as a tacit assumption.[7] I suggest later that this exclusion assumption is central to the problem of agential activity.

Furthermore, Velleman (1992) and Nagel (1986) raise a slightly different version of the agential activity problem. Consider the following quote by Nagel (1986, 110–11).

> Actions seem no longer assignable to individual agents as sources, but become instead components of the flux of events in the world of which the agent is a part. ... There seems no room for agency in a world of neural impulses, chemical reactions, and bone and muscle movements.

Nagel (1986) suggests that causation on the individual level conflicts with causation on the implementation level ("neural impulses, chemical reactions, and bone and muscle movements"). This contrasts with earlier formulations, which suggest a conflict with the intentional level ("wants and beliefs" as in Goldman, 1976, 81). But there is a common element to these different versions of the agential activity problem. It is that event-causal accounts of agency fail to adequately account for the intuition that the agent is active.[8]

Let us try to state the problem more precisely. Any causal account of agency may entail two claims. The first is that *agents are causes*, or for each action there is an agent causing its result. The second claim is that *events are causes*, or for each action there is a (mental) event causing its result. An agent-causal account of agency accepts the first claim, while an event-causal account of agency accepts the second. An agent-causal account locates causation at the

[7] Consider, for example, Chisholm (1976b, 69): "Sometimes a distinction is made between 'event causation' and 'agent causation' and it has been suggested that there is an unbridgeable gap between the two."

[8] More recently, Steward (2012, 197–247) discusses this Nagelian version of the agential activity problem by considering that causation on either the intentional level or the implementation level may conflict with causation on the individual level.

individual level. In contrast, an event-causal account locates causation at the intentional or implementation level.

The agential activity problem arises when we make two assumptions. First, we take it for granted that agents are causes. Second, we assume that the choice between agent-causal and event-causal accounts is exclusive. As a result, any account of agency that focusses on the causal role of events neglects agents.

> *The Agential Activity Problem.*
> **Relevance of Agents.** Intuitively, agents are causes.
> **Exclusion Assumption.** Any account of agency entails either that agents are causes or that events are causes, but not both.
> **Problem.** Any account of agency that entails that events are causes neglects the relevance of agents.

The exclusion assumption is central to the agential activity problem. The entire debate between agent-causal and event-causal accounts of agency seems to be premised on such an exclusion assumption. It is widely assumed that an account is either an event-causal or an agent-causal account of agency, but never both. Whether an account of agency is an event-causal or an agent-causal account depends on what it entails.[9]

I argue that this is a plausible way of thinking about what is at stake in the debate between agent-causal and event-causal accounts of agency. The problem of agential activity is an exclusion problem.[10] If this is true, then we have the chance of moving the debate forward and out of its current stalemate.

[9] There are two alternative ways of distinguishing between agent-causal and event-causal accounts. Both are problematic.

First: Instead of distinguishing accounts by what they entail, we can distinguish them by their definitions or *analyses* of action as follows. While agent-causal accounts analyse actions in terms of agents, event-causal accounts analyse actions in terms of (mental) events (both accounts may entail that agents are causes *and* that actions are causes). But this way of drawing the distinction understates the depth of the disagreement. The literature seems to address a substantive question about different causal claims rather than a question about the terms of description in an analysis.

Second: Some characterise an agent-causal account as the claim that "agent-causation is ontologically irreducible" (2012, 106, and similarly Chisholm, 1976a, 199). This conflates the issue of whether agents are causes, with the ontological issue of whether their causation is reducible. These are distinct issues (cf. Steward, 2012, 203).

[10] This claim itself is not novel. But neither has the analogy been demonstrated, nor has it been made precise. The contribution of this chapter is to rectify this shortcoming. The notion that the problem of agential activity is an exclusion problem has been suggested by Buckareff (2012, 115) and is assumed by Steward (2012, 197–247).

In the next section I explain more precisely what an exclusion problem is. Furthermore, I clarify how we should understand the metaphor of there being different levels (the individual, intentional, and implementation levels).

7.2 Exclusion Problems

A different but well-known exclusion problem arises for certain views in philosophy of mind. Specifically, it arises for positions claiming that behaviour is caused by intentions. This problem is similar to the problem of agential activity insofar as both concern causation on adjacent levels. However, in contrast to the problem of agential activity it concerns the lower two of our three levels. While the agential activity problem concerns the individual and the intentional level, the exclusion problem in philosophy of mind concerns the intentional and the implementation level.[11] Before turning to the details of the problem, I first explain the metaphor of levels. What do we mean when we speak of levels?

By a *level* we understand a way of describing the state of the world.[12] We understand the relative positioning of levels in terms of supervenience.[13] Higher-level states supervene on lower-level states. Supervenience is an asymmetric covariance relation (cf. McLaughlin and Bennett, 2011). A higher-level state supervenes on a lower-level state if and only if there cannot be a change to the higher-level state without a change in the lower-level state. Each level is occupied by a set of entities. One entity is a higher-level entity than another if and only if the level that the first entity occupies supervenes on the level that the second entity occupies.

The exclusion problem arises for certain theories in philosophy of mind claiming that intentions cause behaviour. According to non-reductive physicalism, the mental does not reduce to the physical; mental states, mental events, and

[11] The exception are Nagel and Vellemann, who construe the agential activity problem as a conflict between the individual and the implementation level. But if my reconstruction is correct, then at the heart of the problem is a concern about concurring causes on different levels, irrespective on which levels the causes occur.

[12] I am using the formalism of List and Pivato (2015). See Appendix B. Let S be the set of all states the world may be in. Formally, a level corresponds to a partition of S into equivalence classes.

[13] I stipulate that the relevant relation in the level metaphor ("s is higher-level than b") is the supervenience relation. This should be uncontroversial unless it is assumed that the supervenience relation must be reflexive (everything supervenes on itself). This would conflict with the relation between levels, which needs to be asymmetric. That is, for any s and b, if s is higher-level than b, then b is not higher-level than s.

mental properties are distinct from neural states.[14] Nevertheless, intentions supervene on neural states because there cannot be a change in what someone intends without a change in her neural states (that is, the properties of her neurophysiological system). These claims form the basis of non-reductive physicalism. They lead to a contradiction when we assume the following exclusion principle based on the idea that no effect is caused twice.[15]

> *Exclusion Principle.*
> For all distinct s and s^* such that s^* is higher-level than s, s and s^* do not both cause x.

Non-reductive physicalism seems to neglect the causal relevance of neural properties. Its opponents argue that people's movements are caused by properties of their brain's motor cortex. In this sense, activity within their neurophysiological system is sufficient to cause behaviour. The exclusion principle rules out that the cause of the behaviour can rest on both intentional and implementation levels.

> *Exclusion Problem for Non-reductive Physicalism.*
> **Neural properties are causes.** Neural properties cause behaviour.
> **Supervenience.** Intentions are higher-level than neural properties and are distinct from them.
> **Intentions are causes.** Intentions cause behaviour.
> **Problem.** This is inconsistent with the exclusion principle.

Non-reductive physicalism becomes inconsistent in the face of this exclusion problem. If we accept the exclusion principle, then at most two of the three assumptions underlying non-reductive physicalism can hold. There are several responses to this line of reasoning. I briefly outline four options, each corresponding to one of the assumptions in the exclusion argument.

The first option is to deny the exclusion principle. Upon closer inspection the principle might be false, as argued, for example, by Stoljar (2010, 215–20).

[14] The ontological categories do not matter here. The relevant view in philosophy of mind can be formulated for states, events, and properties respectively.

[15] This idea is more naturally captured by Kim's (1989; 1998; 2005) formulation in terms of sufficient causes, suggesting that the principle is related to a theory of causation as production (List and Menzies, 2009, 489). I assume a general version of the exclusion principle to keep my observations in this chapter neutral about the ontological kind under discussion. Depending on the ontological kind of interest, there are different ways of refining this principle. For a property version, we prefix "for all properties s and s^*", for a state version, we prefix "for all states s and s^*", and similarly for an event or substance version.

The second option is to deny that intentions are causes, while holding on to the supervenience claim. Intentions supervene on neural properties, but they are merely phenomena that are devoid of a causal role. This is the epiphenomenalist position. Another version of this second option draws a distinction. It suggests that intentions are causes in a different sense than the sense in which neural properties are causes. There might be a difference in what we mean by "cause" depending on whether we talk about intentions or neural properties. Jackson and Pettit (1990, 115) draw such a distinction between *causal efficacy* and *causal programming*. In their view, while neural properties are causally efficacious for the effect, intentions causally programme it.

The third option is to deny the supervenience claim. The result is that intentions and neural properties occupy the same level. Moreover, intentions and neural properties are identical rather than distinct entities. So-called mind–brain identity theories comprise a family of views that qualify as this third option.

The fourth option is to deny that neural properties are causes. This option is put forward, for example, by List and Menzies (2009). The resulting view is that the exclusion principle is true, that intentions occupy a higher level than neural properties and are distinct from them, and that intentions cause behaviour but neural properties do not. Neural properties produce behaviour but they do not cause it.

7.3 Agential Activity as an Exclusion Problem

Now that I have detailed the exclusion problem in philosophy of mind, I aim to establish an analogy between it and the problem of agential activity. In this section I describe the analogy, identify challenges in its formulation, and give an overview of the different ways of avoiding the agential activity problem. Instead of defending any particular option, I only point to a path forward for resolving it by framing the problem of agential activity as an exclusion problem.

I make two assumptions explicit that have been implicit so far. First, I assume that agents are distinct from their intentions.[16] I do not deny that intentions may be a part of an agent but I assume that the two are not identical. Second, agents occupy a higher level than intentions. It should be noted that the level

[16] I use "agents" and "individuals" interchangeably here.

of intentions includes various psychological entities including many different mental attitudes, such as intentions, beliefs, and the like. Furthermore, it may also include other events, such as the results of actions.[17]

Proponents of agent-causal accounts of agency make both assumptions. They assume that agents are distinct from their intentions or mental events.[18] Consider, for example, Taylor (1966, 111), who writes that an agent is "never identical with [what] ... is usually proposed as the 'real cause' of [the agent's] act, such as some intention or state of willing". Proponents of agent-causal accounts also plausibly accept the second assumption. Buckareff (2012, 106; and similarly Chisholm, 1976a, 199), for example, defines agent-causation as an entity being "ontologically irreducible" to events. This wording of reducibility suggests that we are talking about entities on different levels.

With these assumptions, the argument behind the problem of agential activity is analogous to the exclusion problem for non-reductive physicalism. We can formulate the agential activity problem as an exclusion problem.

> *Agential Activity as an Exclusion Problem.*
> **Events are causes.** Intentions cause results of actions.
> **Supervenience.** Agents are higher-level than intentions and are distinct from them.
> **Agents are causes.** Agents cause the results of actions.
> **Problem.** This is inconsistent with the exclusion principle.

The structure of this problem is analogous to the structure of the exclusion problem for non-reductive physicalism. Both rest on the exclusion principle. In the domain of philosophy of mind, the exclusion principle creates a problem for non-reductive physicalism, and in philosophy of action it creates a problem of event-causal accounts of causation. Although they use the same principle, they use it in different ways. In philosophy of mind exclusion is invoked to rule out causation on the higher level. In philosophy of action exclusion is invoked to rule out causation on the lower level.[19] When we

[17] In fact, agency as difference-making requires that intentions and results of actions occupy the same level. See Appendix B.

[18] It is important to be clear about the meaning of "distinct". The assumption is that agents are distinct from their intentions in the same way in which intentions are distinct from their neural implementations. In the words of Stoljar (2008), we assume that agents and intentions are numerically and weakly modally distinct, but they are not distinct in a mereological sense. I do not deny that neural implementations can be *a part of* an intention or that an intention can be *a part of* an agent.

[19] The exclusion principle can be used in two different directions, which correspond to the so-called upwards and downwards formulations (List and Menzies, 2009, 490). The argument

add the claim that agents are causes and that they are higher-level entities than intentions, we have an argument for the conclusion that event-causal accounts of agency are false.

> *Agential Activity Argument against Event-causation.*
> **Agents are causes.** Agents cause the results of actions.
> **Supervenience.** Agents are higher-level than intentions and are distinct from them.
> **Exclusion assumption.** If agents are causes, then there is no other thing on a different level that is a cause of the same result of an action.
> **Conclusion.** Events are not causes. Any event-causal account of agency is false.

The main challenge in formulating the problem of agential activity as an exclusion problem concerns the nature of agents and intentions. So far, we have said nothing about what agents and intentions are. What kinds of things do we find on the different levels? The literature varyingly uses the terms "state", "event", and "property" to describe intentions and objects on the implementation level. We might talk, for example, of mental states, mental events, or neural properties. Furthermore, how do individuals fit into this picture of properties and events? For present purposes, these substantive questions can be set aside. The two problems are analogous in virtue of their form rather than their substance, and specifically in virtue of the fact that both rest on the exclusion principle.

Now we return to the options for addressing the agential activity problem. Analogous to the exclusion problem for non-reductive physicalism, we have four options.

The first option is to deny the exclusion principle. When we specify what agents are and how they relate to their intentions, the principle may lose its plausibility. After all, it has always been taken for granted in the discussion between agent-causal and event-causal accounts of agency. If the exclusion principle is false, we do not need to choose. An account of agency may be both an agent-causal and an event-causal account.

The second option is to deny that agents are causes. This is supposedly how some proponents of event-causal accounts of agency respond to the

against non-reductive physicalism uses the upward formulation, while the argument against event-causation uses the downward formulation.

problem of agential activity. After all, or so they argue, causation is a relation between events or between properties, while agents are taken to be substances. Defending the notion that agents are causes would require introducing a new ontological kind into the causal relation. Proponents of event-causal accounts of agency might argue that this fails to make sense of causation in the right way.

The third option is to deny that agents occupy a higher level than their intentions. It is hard to make out the resulting view. Perhaps it is best to understand it as an analogy with the corresponding view in philosophy of mind. The result of denying supervenience might be that agents are identical to their intentions. The disadvantage of such an identity theory would be that it reduces agent-causal accounts of agency to event-causal accounts. This is something that many agent-causalists would want to avoid. But there are alternatives besides crude agent–event identity theories. Perhaps an agent is not identical to all her intentions, but rather an agent is identical to a certain privileged set of intentions that are involved in causing the results of an action. It is in virtue of the contribution of these essential intentions that an agent is active. This view seems to be put forward by Velleman (1992).

Finally, the fourth option is to deny that events are causes. Only agents themselves cause the results of their actions. Some proponents of agent-causal accounts of agency deny the causal relevance of intentions. With the analogy that I have put forward here, proponents of this view now have one additional argument. As an analogy to how List and Menzies (2009) defend non-reductive physicalism, their proposal may be immediately put to work to defend agent-causal accounts of agency.

We can now see that there are several advantages to framing the problem of agential activity as an exclusion problem. It helps us to understand it. In particular, the central role of the exclusion principle becomes clear. This improved understanding grants us a better chance of solving the agential activity problem. Once we clearly define the assumptions underlying the contradiction, we can recognise different options of avoiding it, some of which may have been overlooked. In particular, the first option of denying the exclusion assumption has, to my knowledge, so far received no attention at all.

7.4 Conclusion

In this chapter I have developed an analogy between two problems residing in different areas of philosophy. I have argued that the problem of agential activity is an exclusion problem. The problem of agential activity is premised on the assumption that agents cause the results of their actions. I have argued that there is a further assumption that has often been overlooked. The so-called exclusion assumption requires us to choose between whether results of actions are caused by agents or by their mental states. This conflict parallels the exclusion problem for non-reductive physicalism, which requires us to choose between whether behaviour is caused by mental states or by neural states. Once we spell out this analogy, we may exploit it in our attempts to find solutions to each respective problem.

Chapter 8

Concluding Remarks

What does it take for someone to be responsible for something? I have focussed on one specific aspect of this question and investigated in what sense moral responsibility requires agency. My thesis has advanced the claim that responsibility requires an agent to make a difference. Insofar as the conditions for difference-making closely resemble Lewis' (1973a) conditions for causation, this conception of agency provides causal foundations for moral responsibility while addressing several challenges to existing causal accounts of agency.

The proposal I put forward is particularly apt for cases involving more than one individual. In the introduction I have given two examples. The first involves a vertical group. A commander orders a team of soldiers to shoot an innocent civilian. This case featured prominently in Chapters 4 and 5. I have argued that responsibility in vertical groups is individual responsibility. Specifically, for any action performed on behalf of a vertical group, each individual who reliably gives an order to a subordinate to carry out an action, can be held fully morally responsible for that action, as far as Responsibility Requires Agency is concerned. On each level in the hierarchy of ideal vertical groups, we find individuals who make a difference.

The second example involves a horizontal group, namely an investment committee making a collective decision. I have defended that the group as a whole can be held responsible. This corroborates the words of John Steinbeck's protagonists: "It's not us. It's the monster. The bank isn't like a man." In Chapters 2 and 3 I have argued against alternative answers. Specifically, I argued that collective action exists, in contrast to the view that there is only individual responsibility. I furthermore argued that the

individual merely performs the action on behalf of the group, in opposition to the view that only the person performing the action is responsible.

With the following concluding remarks I look back on the thesis. In Section 8.1 I qualify the proposal by identifying some of its limitations. Section 8.2 contains methodological remarks explaining which higher-order methodological commitments undergird my proposal.

8.1 Limitations

There are three classes of limitations. First, there are *topical limitations* or omissions. I identify some issues that I simply have not discussed. Second, there are *substantive limitations* of the proposal. These are problems or situations that the proposal may fail to address. In particular, I discuss problems in so-called cases of overdetermination. Third, there are *technical limitations*. These are limitations of the proposal arising from the particular choice of formal setup. The formal framework that I have used does not allow for certain distinctions, and imposes some strict requirements. I briefly disclose those.

8.1.1 Topical Limitations

This thesis has pursued limited aims. A focus on specific aims implies the deliberate choice to put some questions aside. In particular, I have omitted two topics. One concerns questions of action ontology and intensionality. Another is the issue of responsibility in horizontal groups.[1]

The first topical limitation concerns two issues that constitute central questions in philosophy of action (cf. Mele, 2012, 369). One involves the nature of actions, or in philosophical jargon, their ontology. What are actions? What kind of things are they? The other limitation concerns the descriptions of actions, or their intensionality. Is the action of flipping the switch the same as the action of turning on the light? Does the agent only intend one thing but not the other? And finally, what does it mean for an agent to do something intentionally? I have left these issues aside because they were not among the

[1] A third topical limitation concerns the problem of so-called deviant causal chains (Davidson, 1980; Smith, 2012). In such cases "the causal pathway between mind and bodily movements deviate [*sic*] from what is normal or expected" (O'Brien, 2014).

proximal questions that came up during my formulation of the main aims of the thesis.

Despite having left them aside, my proposal relates to each of these issues in some way. First, it is neutral on the issue of action ontology and it is compatible with different positions. Moreover, it even provides a simple way of distinguishing between different views on the ontology of actions. I defend these claims about my proposal in Appendix A. Second, on the issue of intensionality, I hypothesise that the current formal framework could be extended to accommodate a formalisation of meaning and descriptions of actions. Finally, I hope that agency as difference-making may provide the formal tools to help to analyse the concept of doing something intentionally.

The second topical limitation concerns the issue of responsibility in horizontal groups. In my thesis I have made room to defend the possibility of collective responsibility, or the argument that a group as a whole can be held responsible. First, I advanced that there are collective actions (Chapter 2) and that individuals may perform actions on behalf of the group agent (Chapter 3). Second, I developed agency as difference-making to explain how, in situations of collective agency, both the group and each individual causally relate to the collective action (Chapters 4 and 5).[2]

Nevertheless, the thesis stops short of discussing several issues. A central question is whether groups are moral agents simpliciter, or in the words of Pettit (2007), whether groups are sufficiently "incorporated". A limitation of this thesis is understanding agency only as a relation, instead of describing groups as moral agents in any non-relational sense. However, moral responsibility for an action requires more than just being an agent of that action (see my discussion in Section 4.1 of Chapter 4). Moreover, the central question for agency in the relational sense is whether groups have intentions. I have appealed briefly to a functionalist analysis of intentions without detailing the full argument (see my discussion of the Discursive Dilemma and of Hive Minds in Chapter 2). In this thesis, questions about the moral agency simpliciter of groups and about moral responsibility in horizontal groups arose only occasionally. Hence, I consider the issue of moral responsibility in horizontal groups to be a second topical limitation, along with remaining

[2]The question is how a group agent on the one hand, and an individual agent on the other hand, can both control the outcome of a collective action. By answering this question, agency as difference-making provides an alternative solution to a problem identified by List and Pettit (2011, 158–63; see also Pettit, 2007; Szigeti, 2014), which arises for positions of collective responsibility.

neutral on central issues of philosophy of action, such as the ontology and intentionality of actions.

8.1.2 Substantive Limitations

An important substantive limitation concerns cases of overdetermination. We have a case of *overdetermination* if there are multiple distinct actual events $c_1, ..., c_n$ and e such that each c_i leads to some effect e. However, the effect e would have occurred without any c_i's individual occurrence (Lewis, 1973a, 567, Schaffer, 2003b, 23).[3] Specifically, we have overdetermination if the positive but not the negative difference-making counterfactual conditional is true for a family of events $c_1, ..., c_n$ with respect to an event e.[4] Cases of overdetermination pose a problem for my account of agency.

> **Overdetermination**. Suppose there are two buttons. If you press either of them, an expensive vase is shattered at a certain time. Suzy and Bill each, unbeknownst to each other, press one of the buttons. The vase is shattered.

There are three actual distinct events. Bill and Suzy individually press their corresponding buttons (c_1 and c_2) and the vase is shattered (e). Suppose that for each c_1 and c_2 the positive difference-making counterfactual is true. If the button were pressed, then the vase would be shattered. But suppose that the negative difference-making counterfactual is false. If either Bill or Suzy were not to press his or her button, then the vase might be shattered anyway because the other might press his or her respective button. For simplicity, let us focus on Bill.

The problem for my proposal is that this seems to be a case of responsibility without agency (cf. Sartorio, 2004, 317–18). Bill seems responsible for shattering the vase. However, agency as difference-making determines that Bill is

[3]Note that I here require overdetermination to involve *actual* events. If we relax this requirement, then we include cases of weak overdetermination, such as the Klingons case in Section 5.3.1.

[4]I have two clarifications. First, my definition of overdetermination is slightly stronger than the orthodox definition. I require that c_i (for all i) and e occur in the actual world *and* that the respective positive difference-making counterfactuals for c_i are true (while the negative ones are false as in any case of overdetermination). The orthodox definition only requires that c_i and e occur in the actual world. The orthodox definition does not make the stronger requirement that the positive counterfactual is true. Second, I name all cases of this sort "overdetermination cases". I neglect the distinction between pre-emption (or asymmetric overdetermination) and overdetermination (or symmetric overdetermination) to keep the exposition simple.

not an agent of shattering the vase. This suggests either that responsibility is not necessary for agency or that agency as difference-making is not an adequate account of agency.

Overdetermination cases may be seen as indicating a substantive limitation of my proposal. I reason that it is rather a technical or a topical limitation. For the first, there are means of amending the formal framework. This thesis relies on a version of Lewis' (1973a, 567) semantics for counterfactuals. Alternatives that can potentially avoid the overdetermination problem include Lewis (2000, 182), Hitchcock (2001), Halpern and Pearl (2005), Woodward (2003, 83–86), and Spohn (2008). There is no quick answer to the question of how exactly agency as difference-making could be formulated in each of these alternative formal frameworks. So far, any one of these could be a viable alternative.

Furthermore, overdetermination cases might indicate a topical limitation. A response may be found by attending to the ontology, and particularly the individuation, of actions.[5] The argument from overdetermination assumes that there is *one* event e, the shattering of the vase, which is the result of Bill pressing his button but also the result of Suzy pressing her button. Suppose we assume that there are *two* distinct events (say, e_1 and e_2); one is the shattering of the vase *by Bill* and the other the shattering of the vase *by Suzy*. For each button pressing there is a different result, depending on who pressed the button. According to this view, for Bill and Suzy there is each an distinct actual event, e_1 or e_2 respectively, to which their pressing their button makes a difference. As a result the argument from overdetermination is deflected.

I have hinted at this option at various points in the thesis, namely in discussing the Klingons case in Section 5.3.1. A similar thought is at play in Section 4.2.2, where I explain a subordinate's low-level autonomy by suggesting that there is a distinct event, the killing *by the subordinate*, over which the subordinate has control. I explore some of these ideas in more detail in Appendix A.

In summary, this suggests that overdetermination cases are not a substantive limitation of the proposal, and reflect instead a technical or topical limitation. We might respond to the argument from overdetermination either by amending or exchanging the formal framework, or by adding further substantive claims to the proposal.

[5] Lewis (1986b, 194–99) hints at a solution in this spirit but dismisses it.

8.1.3 Technical Limitations

Finally, I want to briefly mention two technical limitations. The first limitation is that the formal setup I use here is simple but blunt. It does not allow us to express some of the things we might want to express. The second limitation is that the proposal rules out certain pairs of events as standing in the difference-making relation.

A first technical limitation is that the formal setup does not account for time and space. For the purposes of this thesis, events are sets of worlds. I represent an event as the set of possible worlds in which the event occurs. But, of course, events occur at a location and at a time within a possible world. The worlds I have considered here have no such structure. The atoms of the algebra of my semantics are worlds that are unstructured. This is a technical limitation.

We can extend the proposal to make room for worlds that are structured (see Appendix B). In effect, we would be able to have a more fine-grained representation of events. They would be sets of parts of worlds. For example, events would then be represented as sets of space-time regions at which an event occurs. We might have relations such as "is earlier than" or "is later than" that hold between events. This extension would not be difficult to implement. But it would remain to be seen if it would raise new challenges for my proposal, or if it would give us any substantive advantages.

A second technical limitation concerns so-called backtracking counterfactuals (I draw here on Collins et al., 2004b). Normally, we consider counterfactuals between an earlier and a later event. For example, when we consider whether it is the case that if Suzy were to press her button then the vase would shatter, we take it that pressing the button occurs before the shattering of the vase. The limitation is that we only consider counterfactuals of this sort, namely so-called non-backtracking counterfactuals. For backtracking counterfactuals, this reasoning is temporally reversed. Suppose you fire your gun at a shooting range only if there is nobody in the way. The following counterfactual might be true: if you had fired your gun then there would have been nobody in the way. But this dependence sounds odd and it creates problems. We need to rule it out.[6] This might affect my proposal here. In particular, it might affect using difference-making as a solution for the problem of profligate omissions (Chapter 6). This is a second technical limitation.

[6]There are different ways of implementing the limitation of ruling out backtracking counterfactuals. Crucially, we need to impose restrictions on the similarity ordering (Lewis, 1979).

8.2 A Note on the Method

One of the strengths of agency as difference-making is its simplicity. In this section I illustrate in what ways agency as difference-making is simple by distinguishing between three kinds of conditions for evaluating a philosophical theory. Conditions of the first kind concern how the theory relates to intuitions. Conditions of the second kind concern how the theory relates to linguistic data. Because they relate the theory to some non-theory data, conditions of both of these first two kinds are *external conditions*. Conditions of the third kind are *internal conditions*, such as consistency, axiomatisability, and parsimony. Simplicity falls into this last category.

The structure of this section is as follows. First, I describe how I think about philosophical theories in general. Second, I discuss the three kinds of conditions for evaluating a theory. In this discussion I draw on List and Valentini (2014). Third, I clarify what I mean by simplicity. Specifically, I expand on three internal conditions. I argue that agency as difference-making has the following properties: continuity, minimality, and unity.

8.2.1 Theory Construction

We represent a *theory* as a set of statements, each of which is either true or false. An important subset of a theory are *principles*, which allow us to deduce other statements and judgments. Central principles of moral philosophy are normative principles. They are statements about what an agent ought to do. The utilitarian normative principle says, for example, "you ought to maximise total utility".

According to my view, constructing a theory involves two tasks: optimisation and selection. Optimisation means that any theory should be as good as possible, that is, it should satisfy the conditions with which we evaluate it to the greatest possible extent. For example, we should make sure that a model in economics fits existing empirical data. Fit with existing data is a virtue. But it is only one virtue among many. Another virtue is predictive accuracy, that is, fit with future data. These two virtues may conflict. We must not overfit our model. If the model fits existing data too well, it looses predictive power. On the frontier of theories that satisfy all relevant conditions to the greatest possible extent, we need to make a choice.[7] This is where selection

[7]The notion of the optimality frontier may be taken to entail that rather strong conditions need to be met (for example, that there is an ordering on the set of theory-alternatives and that

comes in. We need to decide how to trade-off between the different virtues of each theory. How much fit with existing data do we sacrifice for predictive accuracy? We make this trade-off between different virtues of a theory based on our views about their relative importance.

We find ourselves in similar situations with philosophical theories. For example, how much simplicity do we sacrifice for good fit with our intuitions? In Section 5.1 I have suggested that different ways out of the naïve trilemma come at different theoretical costs. We have to pay our way out of the contradiction, in the currency of either simplicity, coherence with linguistic data, or coherence with intuitions. This is one instance in which methodological commitments about the relative importance of theoretical virtues determines our theory choice.[8] Another example of such a trade-off is Lewis' (1986a, 134) defence of modal realism.

> Common sense has no absolute authority in philosophy. ... [I]t's a matter of balance and judgement. Some common sense opinions are firmer than others, so the cost of denying common sense opinion differs from one case to the next. And the costs must be set against the gains. ... I acknowledge that my denial of common sense opinion is severe, and I think it is entirely right and proper to count that as a serious cost. How serious is serious enough to be decisive? — That is our central question, yet I don't see how anything can be said about it. *I* still think the price is right, high as it is. Modal realism ought to be accepted as true. The theoretical benefits are worth it.

8.2.2 Theory Evaluation

As we have already seen in the different escape routes from the naïve trilemma in Section 5.1, I distinguish three kinds of conditions to evaluate theories.[9] The conditions of the first kind concern coherence with intuitions. To elicit

this set is sufficiently rich). As a fallback, we have a dominance notion of optimality. A theory is optimal if it is non-dominated, that is, if there is no other theory that does strictly better on one condition and at least as well on all other conditions. We strictly prefer a non-dominated theory to any dominated theory.

[8] I suspect that some disputes between proponents of rival philosophical theories arise in virtue of disagreements about the relative importance of different evaluation criteria. Some find coherence with intuitions more important than simplicity. For others it is the other way around. They will disagree about whether one theory is better than another.

[9] This taxonomy is not exhaustive. For example, some might want philosophical theories about the law to cohere with legal practices. So another family of external conditions concerns the relation between our theory and existing social practices.

intuitions and observe whether they are consistent with a principle, we often use thought experiments. An example that comes to mind of using this method are the so-called trolley cases. Some claim that there is a difference between killing and allowing to die. We elucidate and test this principle by comparing it to our intuitive responses to hypothetical and often stylised cases (Thomson, 1976).[10]

The conditions of the second kind concern linguistic data. We elicit this data by attending to how an expression is commonly used. For example, we might observe that reports about an agent's intentions are canonically formulated with an infinitival clause such as "Pat intends to go to bed early". Some take this to suggest that intentions are not propositional attitudes, that is, what an agent intends is not adequately described by propositions (Ben-Yami, 1997; Brewer, 2006; Grzankowski, 2015). Propositions are expressed only by that-clauses, as for example in "Pat intends *that* she goes to bed early". Since not all intention reports can be expressed appropriately with that-clauses, there are some intentions that do not have only propositional content, or so the argument goes. We have already seen another argument that relies on linguistic data. It is due to Chisholm (1976a, 199) and I have discussed it in Section 7.1. He argues that the truth of "statements [such] as 'Jones killed his uncle'" immediately implies that "there is a certain event or state of affairs that an agent caused or brought about". This is the linguistic argument in favour of agent causation (see also Steward, 2012, 202).

The conditions of the third kind concern the internal properties of a theory or relations between theories. A fairly straightforward internal property with which we evaluate theories is consistency. Different parts of a theory should not contradict each other. It is harder to pin down exactly what it means for a theory to be simple. I suggest that simplicity is a set of such internal properties.

8.2.3 Simplicity

There are three ways in which agency as difference-making is simple. First, it is continuous with theories of causation. Second, it requires only one kind of information. Third, it provides a unified account of agency. I call these

[10] Intuitions are mostly seen only as a guide and not as strict evidence. If we were to take intuitions as evidence, we would reject any part of a theory that conflicts with a given intuitive judgment. More plausibly, we take intuitive judgments as a guide. We require that a theory is in a so-called reflective equilibrium with intuitive judgments. This means that sometimes the theory and sometimes intuitions need to be revised.

properties continuity, minimality, and unity. I briefly expand on each of them in turn.

First, agency as difference-making is continuous with theories of causation. A theory of causation gives necessary and sufficient conditions for the causal relation, that is, for some A to cause some B. Agency as difference-making is continuous with so-called counterfactual theories of causation. Such theories define the causal relation in terms of counterfactual conditionals. Agency as difference-making uses the same, or very similar, conditions to define the agency relation. In particular, my proposal rests largely on the work of List and Menzies (2009) on mental causation, who in turn amend the proposal of Lewis (1973a).

Second, agency as difference-making is minimal. It requires only limited information to decide, for example, whether someone is an agent of something, or whether some event is the result or a mere consequence of an action. I distinguish between four different kinds of information and argue that for most applications, agency as difference-making requires only what I call basic alethic information.

Basic alethic information concerns what is actual and what is possible in any given world. This information is usually expressed in statements like "p is the case", where p is a basic fact. It may involve *whether* or not an agent has a certain intention. What basic alethic information does not involve is the content of an agent's mental states; that is, information about *what* she believes or intends.[11] Basic alethic information is the same kind of information required to determine whether something causes something else. In some discussions I have referred to it as "proximal possibilities". However, beyond this kind of information, except to distinguish between successful and unsuccessful actions, agency as difference-making needs none of the following three kinds of information. This makes it informationally parsimonious.

First, agency as difference-making does not require *epistemic information* about agents. That is, it does not rely on information about what the agent knows or believes. Second, for most applications, it does not require *intentional information* about agents. Such intentional content includes what an agent wants, hopes, or desires. As such the difference-making framework requires

[11] More formally, basic alethic information is expressed using only the set of possible worlds Ω, a similarity relation on this set, and information about which world in this set is the actual world. In contrast to epistemic and intentional information, to express basic alethic information, we need neither an event algebra, nor relations — such as "intends that" or "believes that" — between individuals and propositions.

information about the existence of some intention but not about its content. This intentional information is required only to distinguish successful actions from unsuccessful ones. Third, agency as difference-making does not require *normative information* about agents. It determines who is an agent of an action independently of whether an individual has an obligation to do something. This stands in contrast to the view that any account of agency or causation needs to be normative (Knobe, 2006; Hitchcock and Knobe, 2010).[12]

Finally, agency as difference-making is a unified account of agency. Rosen (2015, 75) suggests that there need to be two relations of responsibility, one for actions and one for their results. In contrast, agency as difference-making is at the same time about actions and results of actions. The distinction between results and actions is often drawn based on how we describe them. Agency as difference-making is defined in terms of propositions. It is thereby independent of names and descriptions and identifies actions (for which we have names) by their results (sets of possible worlds, for which we do not have canonical names).[13]

The second reason agency as difference-making presents a unified account is that it applies to individuals as well as to groups (for an opposing view see Chant, 2010, 30). It suggests that there is no important difference between the agency of individuals and the agency of groups. Both might be agents in the relational sense that is necessary for responsibility. Agency as difference-making is a functional account of agency. Its conditions only require that something has an intention that plays the right role, namely, that makes a difference.

In sum, agency as difference-making is a simple theory that reflects several theoretical virtues. First, it is continuous with other theories of causation. Second, it is minimal with respect to the information it requires. Only basic alethic information is necessary for determining who is an agent of an action and for distinguishing the results of actions from their mere consequences. Third, it is a unified account of agency. It applies to actions as well as to their results, and it applies to individuals as well as to groups and other kinds of agents (anything that has an intention that makes a difference to the occurrence of an action may be an agent).

[12] We need to distinguish normative information from information about what is normal. Agency as difference-making requires information of the latter but not of the former.

[13] I do not deny that there is a distinction between actions and results. But agency as difference-making does not require drawing this distinction. Neither does it resist that this distinction is drawn. In fact, the formal approach presented here might help in distinguishing between actions and results.

Part III

Appendices

Appendix A

Action Individuation

In this appendix I examine the relation between agency as difference-making and theories of action individuation. In Section A.1 I distinguish two modes of considering actions, each of which could be used as a basis for individuating actions. In Section A.2 I argue that agency as difference-making is neutral with respect to action individuation. I employ the aforementioned distinction to sketch two different theories of action individuation within the formal approach to agency.

A.1 Two Modes of Considering Actions

Let us revisit the Coming Home case (Section 5.3.1). I take it for granted that you are an agent of flipping the switch and of turning on the light. The question of action individuation is this: are you doing two things or only one; and more generally, how many actions are there? I argue that the answer to this question depends on whether the actions are considered globally or locally. When they are considered globally, flipping the switch and turning on the light are distinct actions, but when they are considered locally, the two actions may be identical.

How are we to formulate the distinction between considering an action globally and considering it locally? We cannot formulate this distinction in terms of actions $x \in X$ because we have taken this set as a given primitive. We have done so precisely to avoid any assumption about the formal structure or the metaphysical nature of actions. Different theories of action individuation would give us different sets X. So we investigate actions $x \in X$ indirectly

via their results $O(x)$. We make the following mild assumption about action individuation. It states that two actions are distinct if their results are not identical.

Distinct Actions.
For any $x_1, x_2 \in X$, x_1 is distinct from x_2 if $O(x_1) \neq O(x_2)$.

When we consider an action *globally* we consider its result relative to *all* possible situations (represented as possible worlds). When we consider an action *locally* we consider its result relative only to relevantly similar situations (taking the notion of similarity as given). By "consider an action" I mean that we intersect the outcome of an action with a set of possible worlds. When we consider an action x globally, we intersect its outcome $O(x)$ with all possible worlds Ω. When we consider an action x locally, we intersect its outcome $O(x)$ with a set of nearby possible worlds T. With this distinction in hand we can revisit our sufficient condition for two actions being distinct.

Global and Local Distinctness.
For any $x_1, x_2 \in X$, x_1 and x_2 are ...
- globally distinct if $O(x_1) \cap \Omega \neq O(x_2) \cap \Omega$.
- locally distinct if $O(x_1) \cap T \neq O(x_2) \cap T$.

Considered globally, flipping the switch and turning on the light are distinct actions because they have different results. In some rooms you turn on the light by clapping your hands. In these situations, the light is turned on without a switch being flipped. Hence, the set of worlds where a turning on of the light occurs is not identical to the set of worlds where a flipping of the switch occurs. The results of these two actions are not the same.

Considered locally, however, the actions have identical results. All nearby worlds in which the switch is flipped are also worlds in which the light is turned on and *vice versa*. Therefore, the two actions might be identical when considered locally. We say cautiously "might be identical" because we lack the appropriate assumptions to say more. We can say that the actions are not distinct as far as the sufficient condition for distinct actions is concerned. But to determine whether the actions are identical, we would need further

assumptions.[1] Let us now consider two questions about this distinction between considering an action globally or locally.

First question: What exactly is the set T of nearby possible worlds? In the semantics that we use here, nearby worlds relative to a world ω are those worlds that are equally similar compared to ω, given a similarity relation over the set of possible worlds Ω. Technically speaking, nearby possible worlds are an indifference-class of the ω-similarity relation. As usual with possible world semantics, we take the notion of similarity as given. How similar do worlds need to be to count as nearby? I leave this open. There will be a point at which actions with different results globally considered have the same results locally considered.[2]

Second question: What is this distinction about? Formally, the distinction is clear. But the question is: what is the philosophical interpretation of this distinction between considering actions globally and considering them locally? The distinction might suggest that the words "same" or "identical" change their meaning depending on the context in which they appear (cf. Kratzer, 1977). In one context it might be true to say that "the actions of flipping the switch and turning on the light are the same". But in another context this statement might be false. This proposal appears deflationary because it lends itself to explaining away disagreements by identifying them as verbal disputes. Two parties might disagree whether actions are identical partly in virtue of a disagreement about the meaning of "identical" (cf. Chalmers, 2011).

A.2 Neutrality

Agency as difference-making is compatible with different theories about action individuation. To illustrate this point, I consider an objection claiming that there is no distinction between considering actions globally and locally.

[1] We could use the following assumption. For any $x_1, x_2 \in X$, x_1 is locally identical to x_2 if $O(x_1) \cap T = O(x_2) \cap T$. For reasons having to do with the intensionality of actions, this assumption might be too strong. It should rather be interpreted as a sufficient condition for non-distinctness or indiscernibility.

[2] Consider two actions $x_1, x_2 \in X$ and let us assume that a is an agent of these actions. Agency as difference-making entails that the counterfactuals $I(a, y) \boxdot\!\!\rightarrow O(x_1)$ and $I(a, y) \boxdot\!\!\rightarrow O(x_2)$ are true. Therefore, there is a set of nearby worlds (the smallest $I(a, y)$-permitting sphere), in which each world is a $O(x_1)$-and-$O(x_2)$-world. This is entailed by the assumptions about the semantics and the truth conditions of the counterfactual conditional.

Objection: There is no difference between considering an action globally and considering it locally other than the well-known distinction between action types and action tokens. On a global consideration we have action types, on a local consideration we have action tokens. Since the agency relation holds between individuals and action tokens, X must contain only action tokens. While action types are exemplified across different possible worlds, action tokens are concrete particulars, that is, they occur at exactly one possible world. Across two different possible worlds we cannot find the same action token. The action token of turning on the light in the world in which I clap my hands is distinct from the action token of turning on the light in the world in which I flip the switch. Put formally, for all $x \in X$ the result $O(x)$ is a singleton.

Response: The objection implicitly rests on a particular theory of action individuation.[3] I will argue that agency as difference-making is compatible with this theory of action individuation. In a second step, I will argue that there is an alternative theory of action individuation and that agency as difference-making is compatible also with this alternative theory. This suggests that agency as difference-making is neutral with respect to different theories of action individuation.

Agency as difference-making is compatible with the theory of action individuation as it is outlined in the objection above. We need to make some adjustments to accommodate this theory. In particular, if all $O(x)$ are singletons, we need to adjust the positive difference-making conditional which requires that in all nearby possible worlds in which an agent has some intention, the action occurs. However if an action occurs only at exactly *one* world, then this condition will always be false. To fix this problem, there are two possible solutions. First, we could abandon the weak centring condition. Second, we could reformulate the conditional in terms of resulting action types. With the first adjustment we would lose many of the advantages of agency as difference-making. In particular, there would be no distinction between an action being actually successful and it being robustly successful. Furthermore, it would be difficult to account for omissions. With the second adjustment we would introduce a new primitive notion into our model in order to associate action tokens with their respective types. However, the second adjustment would preserve more advantages of agency as difference-making. On both

[3] I interpret Davidson (1971) as defending such a concrete-particular theory of action individuation. It should be noted that this theory implies strengthening the notion of local consideration to its limit. It suggests considering all actions *actulocally*, that is, by intersecting $O(x)$ with the singleton set $\omega \in \Omega$ containing only the actual world.

ways, however, agency as difference-making can accommodate the theory of action individuation as it is outlined in the objection above.

There is an alternative theory of action individuation, with which agency as difference-making is also compatible. On this theory, while the same action may occur at two different possible worlds, the actions $x \in X$ are action tokens nonetheless. For example, *a's turning on the light* may be one token action on this alternative theory; *b's turning on the light* may be another ($a, b \in A$). Each is a distinct element in X. These two action tokens are of the same action type; each is a turning on of the light (action types are represented as subsets of X).

Let us return to the Coming Home case and consider the action of a's flipping the switch. On this alternative theory of action individuation, a's turning on the light occurs together with the action token of a's flipping the switch. But there are also worlds in which the action token of a's turning on the light occurs and the action token of a's flipping the switch does not occur. Instead, the token action of a clapping her hand occurs together with a's turning on the light. Each of these are action tokens that contrast with an action type, which is the set of all actions of flipping the switch by different individuals.

On this alternative theory of action individuation, the distinction between different modes of considering actions does not collapse into the distinction between action types and action tokens. Considered globally, the action tokens of a's flipping the switch and a's turning on the light are distinct. There are some possible worlds where the one action token occurs but not the other. Considered locally, the two action tokens may be the same because they both co-occur in all nearby possible worlds. But irrespective of whether we consider them globally or locally, both are action tokens of different types.

Appendix B

More Structure

In this appendix I give possible worlds more structure. So far, I have taken them to be unstructured primitives without space, time, or levels of description. Now I represent the entities needed to put agency as difference-making to work in an enriched formal structure. Agency as difference-making relies on intentions, events, and individuals. In addition to clarifying the formal representation of these kinds of entities, I also have some broader aims.

More generally, I have three aims: clarification, generalisation, and reconciliation. First, I offer a clarification. In addition to introducing a formal representation of the entities on which agency as difference-making relies, I also clarify what I mean by "levels". I have appealed to this notion in this thesis (most prominently in Chapter 7). For example, I have assumed that there is the level of neurophysiological events, and that there is the level of mental events, and that the latter supervenes on the former. In this appendix, I clarify this metaphor.

Second, I offer a generalisation. To keep things simple, I have used a rather coarse framework throughout most parts of this thesis. This framework has limitations as I have noted in the conclusion (Section 8.1.3). In this appendix, I put forward a proposal to lift some of these limitations.

Third, I offer a reconciliation. Some find the suggestion that there are different levels suspicious. I argue that this proposal has fewer problematic implications than is usually thought. Consider the following quote by John Heil (2004, 222).[1]

[1] Heil (2005) elaborates this critique at length.

> [W]e should do well to dispense with the voguish 'layered' conception of the world. It is one thing to accept the platitude that reality can be variously described, and then to notice that our descriptions can be ordered in a loose hierarchy. It is another matter to reify the hierarchy, imagining that it maps ontological strata.

To address such concerns, I offer a reconciliation consisting of two points. First, I argue that talk of levels does not "reify the hierarchy". Levels are not entities. There is a distinction between levels and the entities that occupy them. The formal framework helps to make this distinction clear. My proposal seems compatible with the platitude Heil mentions. Levels are descriptions of reality. Second, I argue that talk of levels is compatible with a "loose hierarchy". Levels do not need to be conceived as "a mereological structure, ordered by the part–whole relation" (Schaffer, 2003a, 500), as is usually assumed; neither does talk of levels entail that all levels form a linear hierarchy of supervenience relations. Instead, levels might be cross-cutting and need not supervene on one another.

B.1 Formal Setup

I use the formal setup of List and Pivato (2015) making notational modifications to accommodate any number of levels and not just one higher- and one lower-level. My setup is otherwise largely identical to theirs.[2]

The framework describes a system that evolves over time. Let T represent *time*; it is a set of linearly ordered points. Any state of the system is represented by an element of a suitable *state space* S, which is a set of all possible states of the system. A *history* describes a way the system evolves over time as a temporal path through the state space. Formally, it is a function h from T into S, such that for each point in time t, $h(t)$ is the state of the system at t. As will become clear later, histories play the same role as possible worlds in the simpler setup, which I have used throughout this thesis.

There are different ways of describing the states of the system. Such descriptions are almost always less precise, or more coarse, than what the state space S would permit (I define these notions shortly). Any way of describing the

[2] In contrast to List and Pivato (2015) I explicitly assume that there is a set of equivalence relations to induce different levels. However, this is merely a notational variant of their proposal.

system is associated with a *level*. Formally, a level is a partition of S into equivalence classes. It groups together all states of the system between which a particular way of describing it does not differentiate. Let D (for "descriptions") be the set of equivalence relations $\sim$ on S. For simplicity we assume that D is countable and we write $\sim_n$ for the n^{th} equivalence relation. For each equivalence relation $\sim_n$ there is a level S_n, which is the partition of S into equivalence classes induced by the equivalence relation $\sim_n$ on S. While the elements in D are in no particular order, there is one special case. We reserve $\sim_0$ for the trivial equivalence relation of identity (for any $x, y \in S$, $x \sim_0 y$ if and only if $x = y$). Hence, $S_0 = \{\{s\} : s \in S\}$, which we call the *zero level*. We assume that there is at least this zero level; formally, that D contains at least $\sim_0$. Let S^D be the set of all levels, that is, the set of all partitions of S induced by the equivalence relations in D.

What is the relation between the different levels? Some levels are refinements of others. One level is *finer than* another, if every element of the first is a subset of some element of the second. We write $\leq$ for the refinement relation on S^D. The finer than relation $\leq$ is a partial order.[3] The inverse of the finer than relation is, naturally, the *coarser than* relation. To be clear, we do *not* assume that $\leq$ is a total order.[4] There might be two levels that are not refinements of each other.

If one level S_n is a refinement of another S_m, then each element of the former (the refinement) can be associated with an element of the latter (the coarsening). We denote the function that maps each state in S_n to an associated state in S_m with $\sigma_{n\to m}$. If this relation obtains between two levels, then we say that S_m *supervenes* on S_n. Relatedly, we also say that S_m is *higher-level than* S_n (and conversely, S_n *subvenes on* and is *lower-level than* S_m).[5]

There are also histories on higher levels. Just as the zero-level history is a temporal path through the zero-level state space S_0, so a n^{th} level history is a temporal path through the n^{th} level state space S_n. We write $h_{\langle n\rangle}$ for histories on the n^{th} level state space S_n. We obtain these higher-level histories with the σ-function which applies not just to states but also to histories. For each

[3] That is, the relation is reflexive, antisymmetric, and transitive.

[4] That is, we do not assume that $\leq$ is complete; it is *not* the case that for all $S_n, S_k \in S^D$ either $S_n \leq S_k$ or $S_k \leq S_n$.

[5] It should be noted that being a refinement is a necessary but not a sufficient condition for supervenience. Supervenience may not be coextensive with the higher-level than relation. This is because every level may supervene on itself. The higher-level than relation is asymmetric, while the supervenience relation need not be.

zero-level history $h_{\langle 0 \rangle} \in \Omega_0$, the corresponding higher-level history $h_{\langle n \rangle}$ is the function from T into S_n such that

$$h(t)_{\langle n \rangle} = \underset{0 \to n}{\sigma}\left(h(t)_{\langle 0 \rangle}\right).$$

Likewise for each possible history on each level. We obtain the set of n^{th} level histories Ω_n as the projection of the zero-level histories Ω_0 under the respective σ-function. Formally,

$$\Omega_n = \underset{0 \to n}{\sigma}\left(\Omega_0\right).$$

These sets of possible histories on each level now play the role of sets of possible worlds. To employ them in semantics, we need to associate these sets with event algebras. As before, we define an event as a subset of Ω_n. An *event algebra* is a collection of events that is closed under the set theoretic operations of union ($\bigcup$), intersection ($\bigcap$), and complementation ($\setminus$).

For each level there are entities existing on this level. We denote $\mathcal{E}(\Omega_n)$ as the set of entities on level S_n. What entities exist on a level depends on a given way of describing the system. For example, a psychological way of describing an agent will stipulate that the agent has certain beliefs, fears, or hopes. In contrast, a chemical description of the agent will be very different. For our purposes here we take $\mathcal{E}(\Omega_n)$ as a primitive.

It is useful to keep the sets of entities distinct from event algebras. Otherwise, because event algebras are closed under union, the framework would entail the principle of unrestricted composition, that is, the claim that the fusion of any two objects is a further object.

B.2 Applications and Observations

With this setup in place, it is straightforward to represent events, intentions, and individuals. Events are already formally defined. They are sets of possible histories. Yet, events in this formal sense are a much more general representation than what we colloquially mean by "event". In this formal sense of events, also intentions and individuals are events (or sets of events). This generality is an advantage of the formal setup because we can represent the different kinds of entities with only one formal structure.

Events, for example the event that a certain action occurs, are sets of histories at some level. For present purposes we denote events representing the occurrence of an action "actions". Likewise, an intention is an event in this formal sense. It is the event that some agent has a certain intention (see Section 5.2). Finally, also individuals are represented as events in the formal sense, namely, as the event that there is such and such individual.[6]

I make two assumptions. First, I assume that actions and intentions occupy the same level. Second, I assume that the level of individuals is higher than the level of actions and intentions. The first assumption is necessary because agency as difference-making assumes that actions and intentions are elements of the same event algebra. Otherwise, the counterfactual conditionals in terms of which agency as difference-making is defined would not make sense. The second assumption is motivated by the idea that explanations in terms of intentions give a more detailed description than statements in terms of individuals. This assumption is needed for my construal of the problem of agential activity, which is the topic of Chapter 7. I have argued that the problem of agential activity is an exclusion problem, arising from competing causal claims on different levels. This presupposes that individuals can be represented as entities on a higher level than the level which actions and intentions occupy.

Let us now make three observations. The first observation concerns the status of levels; the second concerns the relation between any two levels in S^D; and the third the structure of this relation over the entire set of levels S^D.

First, the notion of levels here is a formal representation of descriptions of the system. Importantly, levels are not entities. Levels are elements of S^D, partitions of the state space, while entities are defined relative to histories through this state space. The setup is a formal representation of "the platitude that reality can be variously described" but without "reifying the hierarchy", contrary to what Heil (2004, 222) suggests.

Second, the relation between two levels — or ways of describing the system — is the finer than relation. If this relation obtains, we may have a σ-function, which is interpreted as a supervenience mapping, relating lower-level states such as neurophysiological events to higher-level states such as intentions. This contrasts with how the relation between levels is usually conceived. Usually, levels are defined in terms of a part–whole relation that obtains between entities (cf. Schaffer, 2003a, 500). This formalisation in terms of the

[6]Of course, this glosses over deep issues concerning personal identity.

finer than relation between levels has different advantages, one of which is described in the next observation.[7]

Third, levels may be cross-cutting. That means, there might be levels that are not refinements of each other. The finer than relation is not complete. While each level is coarser than the zero level, two levels may partition the state space in a cross-cutting way. A level may have elements that are not a subset the elements of another level and *vice versa*. Formally, as List and Pivato (2015) note, this setup "of representing different levels of description in terms of different algebras of events ... permits the existence of an entire *lattice* of 'levels', partially ordered by the coarse-graining relation".

These observations suggest that talk about levels has surprisingly mild implications. It is entirely compatible with assuming that "our descriptions can be ordered in a *loose hierarchy*" (Heil, 2004, 222, my emphasis). Talk about levels need not reify levels, it need not assume that lower-level entities are parts of higher-level entities, and it need not assume a linearly ordered hierarchy of levels.

[7] Construing levels in terms of a mereological part–whole relation entails that the relation is complete (assuming that everything is either a part of some other object or has at least one part).

Bibliography

Adams, Frederick and Aizawa, Kenneth. 2011. *The Bounds of Cognition*. John Wiley & Sons.

Aguilar, Jesús Humberto and Buckareff, Andrei A. 2010. Causal theory of action: Origins and issues. In *Causing Human Actions: New Perspectives on the Causal Theory of Action*, edited by Aguilar, Jesús Humberto and Buckareff, Andrei A., pages 1–26, MIT Press.

Alston, William P. 1958. Ontological commitments. *Philosophical Studies* 9 (1-2):8–17.

Alvarez, Maria. 2013. Agency and two-way powers. *Proceedings of the Aristotelian Society* 113:101–121.

Alvarez, Maria and Hyman, John. 1998. Agents and their actions. *Philosophy* 73 (2):219–245.

Aristotle. 2014. *Nicomachean Ethics*. Hackett Publishing.

Armstrong, David Malet. 1968. *A Materialist Theory of the Mind*. Routledge & K. Paul.

Baker, Lynne Rudder. 2000. *Persons and Bodies: A Constitution View*. Cambridge University Press.

Ben-Yami, Hanoch. 1997. Against characterizing mental states as propositional attitudes. *The Philosophical Quarterly* 47 (186):84–89.

Bennett, Jonathan Francis. 1988. *Events and Their Names*. Hackett Publishing.

Block, Ned. 2006. Troubles with functionalism. In *Theories of Mind: An Introductory Reader*, edited by Eckert, Maureen, Rowman & Littlefield.

Botvinick, Matthew and Cohen, Jonathan. 1998. Rubber hands 'feel' touch that eyes see. *Nature* 391 (6669):756–756.

Braham, Matthew and van Hees, Martin. 2009. Degrees of causation. *Erkenntnis* 71 (3):323–344.

———. 2011. Responsibility voids. *The Philosophical Quarterly* 61 (242):6–15.

Bratman, Michael. 2014. *Shared Agency: A Planning Theory of Acting Together*. Oxford University Press.

Brewer, Talbot. 2006. Three dogmas of desire. In *Values and Virtues: Aristotelianism in Contemporary Ethics*, edited by Chappell, Timothy, pages 257–284, Oxford University Press.

Buckareff, Andrei A. 2012. How does agent-causal power work? *The Modern Schoolman* 88 (1/2):105–121.

Carmena, Jose M., Lebedev, Mikhail A., Crist, Roy E., O'Doherty, Joseph E., Santucci, David M., Dimitrov, Dragan F., Patil, Parag G., Henriquez, Craig S., and Nicolelis, Miguel A. L. 2003. Learning to control a brain–machine interface for reaching and grasping by primates. *PLoS Biology* 1 (2):e2.

Chalmers, David. 2011. Verbal disputes. *Philosophical Review* 120 (4):515–566.

Chant, Sara Rachel. 2007. Unintentional collective action. *Philosophical Explorations* 10 (3):245–256.

———. 2010. Two problems of composition in collective action. In *New Waves in Metaphysics*, edited by Hazlett, Allan, pages 27–53, Palgrave Macmillan.

Chisholm, Roderick Milton. 1976a. The agent as cause. In *Action Theory*, edited by Brand, Myles and Walton, Douglas, number 97 in Synthese Library, pages 199–211, Springer Netherlands.

———. 1976b. *Person and Object: A Metaphysical Study*. Open Court Publishing Company.

Chockler, Hana and Halpern, Joseph Y. 2004. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research* 22:93–115.

Clark, Andy. 2004. *Natural-born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. Oxford University Press.

Clark, Andy and Chalmers, David. 1998. The extended mind. *Analysis* 58 (1):7–19.

Clarke, Randolph. 2010. Intentional omissions. *Noûs* 44 (1):158–177.

———. 2012. What is an omission? *Philosophical Issues* 22 (1):127–143.

Coleman, James Samuel. 1994. *Foundations of Social Theory*. Harvard University Press.

Collins, John, Hall, Edward Jonathan, and Paul, Laurie Ann, editors. 2004a. *Causation and Counterfactuals*. MIT Press.

Collins, John, Hall, Edward Jonathan, and Paul, Laurie Ann. 2004b. Counterfactuals and causation: History, problems, and prospects. In *Causation and Counterfactuals*, edited by Collins, John, Hall, Edward Jonathan, and Paul, Laurie Ann, pages 1–58, MIT Press.

Colvin, Eric. 1995. Corporate personality and criminal liability. *Criminal Law Forum* 6 (1):1–44.

Cooper, D. E. 1968. Collective responsibility. *Philosophy* 43 (165):258–268.

Copp, David. 1979. Collective actions and secondary actions. *American Philosophical Quarterly* 16 (3):177–186.

Cordeschi, Roberto. 2013. Automatic decision-making and reliability in robotic systems: Some implications in the case of robot weapons. *AI & Society* 28 (4):431–441.

Danner, Allison Marston and Martinez, Jenny S. 2005. Guilty associations: Joint criminal enterprise, command responsibility, and the development of international criminal law. *California Law Review* pages 75–169.

Davidson, Donald. 1963. Actions, reasons, and causes. *The Journal of Philosophy* 60 (23):685–700.

———. 1967. The logical form of action sentences. In *The Logic of Decision and Action*, edited by Rescher, Nicholas, University of Pittsburgh Press.

———. 1971. Agency. In *Agent, Action, and Reason*, edited by Binkley, Robert, Bronaugh, Richard, and Marras, Ausonio, University of Toronto Press.

———. 1980. Freedom to act. In *Essays on Actions and Events*, pages 59–74, Oxford University Press.

DeFalco, Randle C. 2013. Contextualizing actus reus under article 25(3)(d) of the ICC statute thresholds of contribution. *Journal of International Criminal Justice* 11 (4):715–735.

DeMott, Deborah A. 2014. The fiduciary character of agency and the interpretation of instructions. In *Philosophical Foundations of Fiduciary Law*, edited by Gold, Andrew S. and Miller, Paul B., Philosophical Foundations of Law, chapter 16, pages 321–338, Oxford University Press.

Dempsey, James. 2013. Corporations and non-agential moral responsibility. *Journal of Applied Philosophy* 30 (4):334–350.

Dennett, Daniel Clement. 1981. Where am I? In *Brainstorms: Philosophical Essays on Mind and Psychology*, pages 310–323, MIT Press.

———. 1989. *The Intentional Stance*. MIT Press.

Dowe, Phil. 2010. Proportionality and omissions. *Analysis* 70 (3):446–451.

Ehrsson, H. Henrik. 2007. The experimental induction of out-of-body experiences. *Science* 317 (5841):1048–1048.

Eldar, Shachar. 2013. Indirect co-perpetration. *Criminal Law and Philosophy* pages 1–13.

Elster, Jon. 2007. *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge University Press.

Feinberg, Joel. 1970a. Action and responsibility. In *Doing and Deserving: Essays in the Theory of Responsibility*, chapter 6, pages 119–151, Princeton University Press.

———. 1970b. Collective responsibility. In *Doing and Deserving: Essays in the Theory of Responsibility*, chapter 9, pages 222–251, Princeton University Press.

Fischer, John Martin and Ravizza, Mark. 2000. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.

Frankfurt, Harry G. 1969. Alternate possibilities and moral responsibility. *The Journal of Philosophy* 66 (23):829–839.

French, Peter A. 1979. The corporation as a moral person. *American Philosophical Quarterly* 16 (3):207–215.

Goldman, Alvin I. 1976. *A Theory of Human Action*. Princeton University Press.

Grzankowski, Alex. 2015. Not all attitudes are propositional. *European Journal of Philosophy* 23 (3):374–391.

Haddock, Adrian. 2005. At one with our actions, but at two with our bodies: Hornsby's account of action. *Philosophical Explorations* 8 (2):157–172.

Halpern, Joseph Y. and Pearl, Judea. 2005. Causes and explanations: A structural-model approach. Part I: causes. *The British Journal for the Philosophy of Science* 56 (4):843–887.

Heil, John. 2004. *Philosophy of Mind: A Contemporary Introduction*. Routledge, 2nd ed. edition.

———. 2005. *From an Ontological Point of View*. Clarendon Press.

Hertle, Hans-Hermann and Nooke, Maria. 2009. *Die Todesopfer an der Berliner Mauer 1961-1989: Ein biographisches Handbuch*. Ch. Links Verlag.

Hindriks, Frank. 2009. Corporate responsibility and judgment aggregation. *Economics and Philosophy* 25 (02):161–177.

Hitchcock, Christopher. 2001. The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy* 98 (6):273.

Hitchcock, Christopher and Knobe, Joshua. 2010. Cause and norm. *The Journal of Philosophy* 106 (11):587–612.

Hobbes, Thomas. 1909 [1651]. *Hobbes's Leviathan reprinted from the edition of 1651 with an Essay by the Late W.G. Pogson Smith*. Clarendon Press.

Hofstadter, Douglas R. 2000. *Gödel, Escher, Bach: An Eternal Golden Braid*. Penguin Books.

Holton, Richard. 2009. *Willing, Wanting, Waiting*. Oxford University Press.

Hornsby, Jennifer. 1980. *Actions*. Routledge.

———. 2004. Agency and actions. In *Agency and Action*, edited by Hyman, John and Steward, Helen, Cambridge University Press.

Huebner, Bryce. 2014. *Macrocognition: A Theory of Distributed Minds and Collective Intentionality*. Oxford University Press.

ICC. 2007. International Criminal Court: The Prosecutor v. Thomas Lubanga Dyilo, case no ICC-01/04-01/06-803.

———. 2008. International Criminal Court: Prosecutor v. Katanga and Chui, case no ICC-01/04-01/07-717.

Ifft, Peter J., Shokur, Solaiman, Li, Zheng, Lebedev, Mikhail A., and Nicolelis, Miguel A. L. 2013. A brain–machine interface enables bimanual arm movements in monkeys. *Science Translational Medicine* 5 (210).

Isaacs, Tracy. 2011. *Moral Responsibility in Collective Contexts*. Oxford University Press.

Jackson, Frank. 1980. Ontological commitment and paraphrase. *Philosophy* 55 (213):303–315.

———. 1987. Group morality. In *Metaphysics and Morality: Essays in Honour of J.J.C. Smart*, edited by Pettit, Philip, Sylvan, Richard, and Norman, Jean, pages 91–110, Blackwell.

Jackson, Frank and Pettit, Philip. 1990. Program explanation: A general perspective. *Analysis* 50 (2):107–117.

Jerusalem District Court. 1961. The Attorney General v. Eichmann, case no. 40/61.

Kelsen, Hans. 1949. *General Theory of Law and State*. Harvard University Press.

Kim, Jaegwon. 1989. The myth of nonreductive materialism. *Proceedings and Addresses of the American Philosophical Association* 63 (3):31–47.

———. 1998. *Mind in a Physical World*. MIT Press.

———. 2005. *Physicalism, or Something Near Enough*. Princeton University Press.

Kincaid, Harold. 1986. Reduction, explanation, and individualism. *Philosophy of Science* 53 (4):492–513.

Kment, Boris. 2010. Causation: Determination and difference-making. *Noûs* 44 (1):80–111.

Knobe, Joshua. 2003. Intentional action and side effects in ordinary language. *Analysis* 63 (279):190–194.

———. 2006. The concept of intentional action: A case study in the uses of folk psychology. *Philosophical Studies* 130 (2):203–231.

Kornhauser, Lewis A. and Sager, Lawrence G. 1993. The one and the many: Adjudication in collegial courts. *California Law Review* 81 (1):1–59.

Kraemer, Eric Russert. 1978. Intentional action, chance and control. *Analysis* 38 (3):116–117.

Kratzer, Angelika. 1977. What 'must' and 'can' must and can mean. *Linguistics and Philosophy* 1 (3):337–355.

Kutz, Christopher. 2000. *Complicity: Ethics and Law for a Collective Age*. Cambridge University Press.

Lagnado, David A., Gerstenberg, Tobias, and Zultan, Ro'i. 2013. Causal responsibility and counterfactuals. *Cognitive Science* 37 (6):1036–1073.

Lawson, Brian. 2011. Individual complicity in collective wrongdoing. *Ethical Theory and Moral Practice* 16 (2):227–243.

Lenggenhager, Bigna, Tadi, Tej, Metzinger, Thomas, and Blanke, Olaf. 2007. Video ergo sum: Manipulating bodily self-consciousness. *Science* 317 (5841):1096–1099.

Lewis, David. 1973a. Causation. *The Journal of Philosophy* 70 (17):556.

———. 1973b. *Counterfactuals*. Wiley-Blackwell.

———. 1979. Counterfactual dependence and time's arrow. *Noûs* 13 (4):455.

———. 1986a. *On the Plurality of Worlds*. Basil Blackwell.

———. 1986b. Postscripts to "Causation". In *Philosophical Papers: Volume II*, pages 172–213, Oxford University Press.

———. 2000. Causation as influence. *The Journal of Philosophy* 97 (4):182–197.

Lin, Patrick, Bekey, George, and Abney, Keith. 2008. Autonomous Military Robotics: Risk, Ethics, and Design. Technical report.

List, Christian and Menzies, Peter. 2009. Non-reductive physicalism and the limits of the exclusion principle. *The Journal of Philosophy* 106 (9):475–502.

List, Christian and Pettit, Philip. 2002. Aggregating sets of judgments: An impossibility result. *Economics and Philosophy* 18 (01):89–110.

———. 2006. Group agency and supervenience. *The Southern Journal of Philosophy* 44 (S1):85–105.

———. 2011. *Group Agency: The Possibility, Design, and Status of Corporate Agents.* Oxford University Press.

List, Christian and Pivato, Marcus. 2015. Emergent chance. *The Philosophical Review* 124 (1):119–152.

List, Christian and Valentini, Laura. 2014. Political Theory, manuscript.

Lokhorst, Gert-Jan and van den Hoven, Jeroen. 2011. Responsibility for military robots. In *Robot Ethics*, edited by Patrick, Lin, Keith, Abney, and Bekey, George A., MIT Press.

Ludwig, Kirk. 2007a. Collective intentional behavior from the standpoint of semantics. *Noûs* 41 (3):355–393.

———. 2007b. Foundations of social reality in collective intentional behavior. In *Intentional Acts and Institutional Facts: Essays on John Searle's Social Ontology*, edited by Tsohatzidis, Savas L., Springer.

———. 2014a. The ontology of collective action. In *From Individual to Collective Intentionality: New Essays*, edited by Chant, Sara Rachel, Hindriks, Frank, and Preyer, Gerhard, pages 112–129, Oxford University Press.

———. 2014b. Proxy agency in collective action. *Noûs* 48 (1):75–105.

Mackie, John Leslie. 1977. *Ethics: Inventing Right and Wrong*. Penguin Books.

Manacorda, Stefano and Meloni, Chantal. 2011. Indirect perpetration versus joint criminal enterprise concurring approaches in the practice of international criminal law? *Journal of International Criminal Justice* 9 (1):159–178.

Margolis, Joseph. 1974. War and ideology. In *Philosophy, Morality and International Affairs*, edited by Held, Virginia, Morgenbesser, Sidney, and Nagel, Thomas, pages 246–265.

Massey, Gerald J. 1976. Tom, Dick, and Harry, and all the king's men. *American Philosophical Quarterly* 13 (2):89–107.

Matthias, Andreas. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6 (3):175–183.

May, Larry. 1987. *The Morality of Groups*. University of Notre Dame Press.

———. 1996. *Sharing Responsibility*. University of Chicago Press.

Mayr, Erasmus. 2011. *Understanding Human Agency*. Oxford University Press.

McDowell, John. 1996. *Mind and World*. Harvard University Press, 2nd edition.

McGrath, Sarah. 2005. Causation by omission: A dilemma. *Philosophical Studies* 123 (1-2):125–148.

McLaughlin, Brian and Bennett, Karen. 2011. Supervenience. In *The Stanford Encyclopedia of Philosophy*, edited by Zalta, Edward N., winter 2011 edition.

Melden, A. I. 1956. Action. *The Philosophical Review* 65 (4):523–541.

———. 1961. *Free Action*. Taylor & Francis.

Mele, Alfred R. 1992. *Springs of Action: Understanding Intentional Behavior*. Oxford University Press.

———. 2003. *Motivation and Agency*. Oxford University Press.

———. 2012. Intentional, unintenional, or neither? Middle ground in theory and practice. *American Philosophical Quarterly* 49 (4):369–379.

Menzies, Peter. 2004. Difference-making in context. In *Causation and Counterfactuals*, edited by Collins, John, Hall, Edward Jonathan, and Paul, Laurie Ann, pages 139–180, MIT Press.

Metzinger, Thomas. 2006. Reply to Gallagher: Different conceptions of embodiment. *Psyche* 12 (4).

Miller, Seumas. 2001. *Social Action: A Teleological Account*. Cambridge University Press.

Moore, Michael S. 2010. Renewed questions about the causal theory of action. In *Causing Human Actions: New Perspectives on the Causal Theory of Action*, edited by Aguilar, Jesús Humberto and Buckareff, Andrei A., pages 27–44, MIT Press.

Nagel, Thomas. 1986. *The View From Nowhere*. Oxford University Press.

Nair, Prashant. 2013. Brain–machine interface. *Proceedings of the National Academy of Sciences* 110 (46):18343–18343.

Nehring, Klaus. 2005. The (im)possibility of a paretian rational. *Economics working papers, Institute for Advanced Study, School of Social Science* .

Nicolelis, Miguel A. L. and Lebedev, Mikhail A. 2009. Principles of neural ensemble physiology underlying the operation of brain–machine interfaces. *Nature Reviews Neuroscience* 10 (7):530–540.

Nolan, Daniel. 2014a. Hyperintensional metaphysics. *Philosophical Studies* 171 (1):149–160.

———. 2014b. The question of moral ontology. *Philosophical Perspectives* 28 (1):201–221.

O'Brien, Lilian. 2014. *Philosophy of Action*. Palgrave Macmillan.

O'Connor, Timothy. 2005. Freedom with a human face. *Midwest Studies In Philosophy* 29 (1):207–227.

Pagallo, Ugo. 2011. Killers, fridges, and slaves: a legal journey in robotics. *AI & Society* 26 (4):347–354.

Pereboom, Derk. 2014a. The disappearing agent objection to event-causal libertarianism. *Philosophical Studies* 169 (1):59–69.

———. 2014b. *Free Will, Agency, and Meaning in Life*. Oxford University Press.

Pettit, Philip. 2001. *A Theory of Freedom: From the Psychology to the Politics of Agency*. Oxford University Press.

———. 2007. Responsibility incorporated. *Ethics* 117 (2):171–201.

Pettit, Philip and Schweikard, David. 2006. Joint actions and group agents. *Philosophy of the Social Sciences* 36 (1):18 –39.

Pomes, Ausias and Slater, Mel. 2013. Drift and ownership toward a distant virtual body. *Frontiers in Human Neuroscience* 7.

Quine, Willard Van Orman. 1960. *Word and Object*. MIT Press.

Quinton, Anthony. 1975. The presidential address: Social objects. *Proceedings of the Aristotelian Society* 76:1–viii.

Reel, Adolf Frank. 1949. *The Case of General Yamashita*. University of Chicago Press.

Rosen, Gideon. 2015. The alethic conception of moral responsibility. In *The Nature of Moral Responsibility: New Essays*, edited by Clarke, Randolph,

McKenna, Michael, and Smith, Angela M., pages 65–88, Oxford University Press.

Roxin, Claus. 1963. *Täterschaft und Tatherrschaft*. Cram, de Gruyter & Co.

———. 2011. Crimes as part of organized power structures. *Journal of International Criminal Justice* 9 (1):193–205.

Sartorio, Carolina. 2004. How to be responsible for something without causing it. *Philosophical Perspectives* 18 (1):315–336.

———. 2007. Causation and responsibility. *Philosophy Compass* 2 (5):749–765.

———. 2009. Omissions and causalism. *Noûs* 43 (3):513–530.

Sawyer, R. Keith. 2001. Emergence in sociology: Contemporary philosophy of mind and some implications for sociological theory. *American Journal of Sociology* 107 (3):551–585.

Schaffer, Jonathan. 2003a. Is there a fundamental level? *Noûs* 37 (3):498–517.

———. 2003b. Overdetermining causes. *Philosophical Studies* 114 (1-2):23–45.

Schulzke, Marcus. 2013. Autonomous weapons and distributed responsibility. *Philosophy & Technology* 26 (2):203–219.

Searle, John R. 1979. What is an intentional state? *Mind* 88 (349):74–92.

———. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3 (03):417–424.

———. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

———. 1984. *Minds, Brains, and Science*. Harvard University Press.

———. 2010. *Making the Social World*. Oxford University Press.

Seeley, Thomas D. 2010. *Honeybee Democracy*. Princeton University Press.

Shapiro, Scott J. 2011. *Legality*. Harvard University Press.

———. 2014. Massively shared agency. In *Rational and Social Agency: The Philosophy of Michael Bratman*, edited by Vargas, Manuel and Yaffe, Gideon, pages 257–293, Oxford University Press.

Shepherd, Joshua. 2015. Deciding as intentional action: Control over decisions. *Australasian Journal of Philosophy* 93 (2):335–351.

Smidt, Michael L. 2000. Yamashita, Medina, and beyond: Command responsibility in contemporary military operations. *Military Law Review* 164:155–234.

Smith, Michael. 2012. Four objections to the standard story of action (and four replies). *Philosophical Issues* 22 (1):387–401.

von Solodkoff, Tatjana. 2014. Paraphrase strategies in metaphysics. *Philosophy Compass* 9 (8):570–582.

Sparrow, Robert. 2007. Killer robots. *Journal of Applied Philosophy* 24 (1):62–77.

Spohn, Wolfgang. 2008. Causation: An alternative. In *Causation, Coherence and Concepts: A Collection of Essays*, pages 75–98, Springer.

Steinbeck, John. 1939. *The Grapes of Wrath*. Penguin Books.

Steinhoff, Uwe. 2013. Killing them safely: Extreme asymmetry and its discontents. In *Killing by Remote Control: The Ethics of an Unmanned Military*, edited by Strawser, Bradley Jay, Oxford University Press.

Steward, Helen. 2012. *A Metaphysics for Freedom*. Oxford University Press.

Stoljar, Daniel. 2008. Distinctions in distinction. In *Being Reduced: New Essays on Reduction, Explanation, and Causation*, edited by Hohwy, Jakob and Kallestrup, Jesper, Oxford University Press.

———. 2010. *Physicalism*. Routledge.

Sylvan, Kurt L. 2012. How to be a redundant realist. *Episteme* 9 (03):271–282.

Szigeti, András. 2014. Collective responsibility and group-control. In *Rethinking the Individualism–Holism Debate*, edited by Zahle, Julie and Collin, Finn, number 372 in Synthese Library, pages 97–116, Springer International Publishing.

Taylor, Richard. 1966. *Action and Purpose*.

Thomson, Judith Jarvis. 1976. Killing, letting die, and the trolley problem. *The Monist* 59 (2).

Tollefsen, Deborah. 2002. Collective intentionality and the social sciences. *Philosophy of the Social Sciences* 32 (1):25–50.

Tuomela, Raimo. 1989. Actions by collectives. *Philosophical Perspectives* 3:471–496.

———. 1995. *The Importance of Us: A Philosophical Study of Basic Social Notions*. Stanford University Press.

US Department of the Army. 2014. Army Regulation 600–20: Army Command Policy.

Velasquez, Manuel G. 1983. Why corporations are not morally responsible for anything they do. *Business & Professional Ethics Journal* 2 (3):1–18.

Velleman, J. David. 1992. What happens when someone acts? *Mind* 101 (403):461–481.

Wall, Edmund. 2000. The problem of group agency. *The Philosophical Forum* 31:187–197.

Walzer, Michael. 2004. Two kinds of military responsibility. In *Arguing About War*, chapter 2, pages 22–32, Yale University Press.

Wasserstrom, Richard A. 1980. Conduct and responsibility in war. In *Philosophy and Social Issues: Five Studies*, University of Notre Dame Press.

Watkins, John. 1952. The principle of methodological individualism. *The British Journal for the Philosophy of Science* 3 (10):186–189.

———. 1973. Historical explanation in the social sciences. In *Modes of Individualism and Collectivism*, edited by O'Neill, John, Heinemann.

Weigend, Thomas. 2011. Perpetration through an organization the unexpected career of a german legal concept. *Journal of International Criminal Justice* 9 (1):91–111.

Whitney, Courtney. 1950. *The Case of General Yamashita: A Memorandum*. Tokyo: General Headquarters Supreme Commander of the Allied Forces.

Williams, Bernard. 1994. *Shame and Necessity*. University of California Press.

Wilson, Robert A. 2001. Group-level cognition. *Philosophy of Science* pages S262–S273.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Ylikoski, Petri. 2012. Micro, macro, mechanisms. In *The Oxford Handbook of Philosophy of Social Science*, edited by Kincaid, Harold, pages 21–45, Oxford University Press.