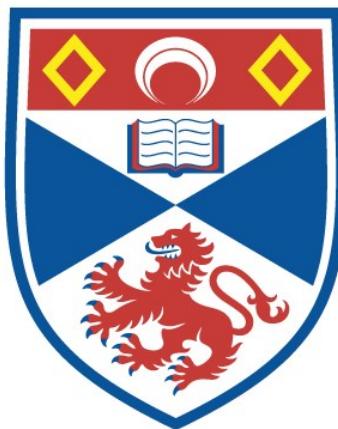


THE NEXUS OF CONTROL: INTENTIONAL ACTIVITY AND  
MORAL ACCOUNTABILITY

Niël Conradie

A Thesis Submitted for the Degree of PhD  
at the  
University of St Andrews



2018

Full metadata for this thesis is available in  
St Andrews Research Repository  
at:  
<http://research-repository.st-andrews.ac.uk/>

Please use this identifier to cite or link to this thesis:  
<http://hdl.handle.net/10023/13660>

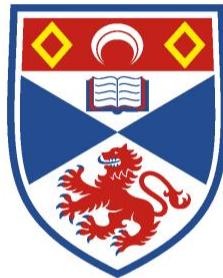
This item is protected by original copyright

This item is licensed under a  
Creative Commons Licence

<http://creativecommons.org/licenses/by-nc-nd/4.0/>

# The nexus of control: intentional activity and moral accountability

Niël Conradie



University of  
St Andrews

This thesis is submitted in partial fulfilment for the degree of PhD  
at the  
University of St Andrews

26<sup>th</sup> January 2018

## **Abstract**

My aim in this thesis is to untangle a conceptual knot at the intersection of moral responsibility and action theory. This knot can be expressed as the following question: What is the relationship between an agent's openness to moral responsibility and the intentional status of her behaviour? My answer to this question is developed in three steps. I first develop a control-backed account of intentional agency, one that borrows vital insights from the cognitive sciences – in the form of *Dual Process Theory* – in understanding the control condition central to the account, and demonstrate that this account fares at least as well as its rivals in the field. This control condition will be explained as the requirement for a kind of oversight over behaviour, namely: *System 2 Oversight*. Secondly, I investigate the dominant positions in the discussion surrounding the role of control in moral responsibility. After consideration of some shortcomings of these positions – especially the inability to properly account for so-called *ambivalence cases* – I defend an alternative *pluralist* account of moral responsibility, in which there are two co-extant variants of such responsibility: attributability and accountability. The latter of these will be shown to have a necessary control condition, also best understood in terms of a requirement for oversight (rather than conscious or online control), and in terms of the workings of the dual system mechanism. I then demonstrate how these two accounts are necessarily related through the shared role of this kind of control, leading to my answer to the original question: *if an agent is open to moral accountability based on some activity or outcome, this activity or outcome must necessarily have positive intentional status*. I then apply this answer in a consideration of certain cases of the use of the Doctrine of Double Effect.

## **1. Candidate's declarations:**

I, **Niël Conradie**, hereby certify that this thesis, which is approximately **80000** words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in **January 2011** and as a candidate for the degree of **Doctor of Philosophy** in **November 2014**; the higher study for which this is a record was carried out in the University of St Andrews between **2014** and **2018**.

(If you received assistance in writing from anyone other than your supervisor/s):

I, ..... received assistance in the writing of this thesis in respect of [language, grammar, spelling or syntax], which was provided by .....

Date **23/01/2018** signature of candidate .....

## **2. Supervisor's declaration:**

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of **Doctor in Philosophy** in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date **23/01/2018** signature of supervisor .....

## **3. Permission for publication: (to be signed by both candidate and supervisor)**

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

### **PRINTED COPY**

- (a) No embargo on print copy  
(b) Embargo on all or part of print copy for a period of ... years (maximum five) on the following ground(s):
  - Publication would be commercially damaging to the researcher, or to the supervisor, or the University
  - Publication would preclude future publication
  - Publication would be in breach of laws or ethics

(c) Permanent or longer term embargo on all or part of print copy for a period of ... years (the request will be referred to the Pro-Provost and permission will be granted only in exceptional circumstances).

### **Supporting statement for printed embargo request if greater than 2 years:**

### **ELECTRONIC COPY**

- (a) No embargo on electronic copy  
(b) Embargo on all or part of electronic copy for a period of ... years (maximum five) on the following ground(s):
  - Publication would be commercially damaging to the researcher, or to the supervisor, or the University
  - Publication would preclude future publication
  - Publication would be in breach of law or ethics

(c) Permanent or longer term embargo on all or part of electronic copy for a period of ... years (the request will be referred to the Pro-Provost and permission will be granted only in exceptional circumstances).

### **Supporting statement for electronic embargo request if greater than 2 years:**

### **ABSTRACT AND TITLE EMBARGOES**

An embargo on the full text copy of your thesis in the electronic and printed formats will be granted automatically in the first instance. This embargo includes the abstract and title except that the title will be used in the graduation booklet.

If you have selected an embargo option indicate below if you wish to allow the thesis abstract and/or title to be published. If you do not complete the section below the title and abstract will remain embargoed along with the text of the thesis.

- a) I agree to the title and abstract being published
- b) I require an embargo on abstract
- c) I require an embargo on title

YES/NO  
YES   
NO   
YES   
NO

Date **23/01/2018** signature of candidate .....

signature of supervisor .....

*Please note initial embargos can be requested for a maximum of five years. An embargo on a thesis submitted to the Faculty of Science or Medicine is rarely granted for more than two years in the first instance, without good justification. The Library will not lift an embargo before confirming with the student and supervisor that they do not intend to request a continuation. In the absence of an agreed response from both student and supervisor, the Head of School will be consulted. Please note that the total period of an embargo, including any continuation, is not expected to exceed ten years.*  
*Where part of a thesis is to be embargoed, please specify the part and the reason.*

## **Acknowledgements**

Thanks to my folks and my brothers, whose support was ever-present. Thank you to Justin, without whom this work would not exist, and to Professors Broadie and Gaut, whose insightful comments have helped shape it. And loving thanks to Rina, for being an inestimable inspiration, and a constant guiding star in the firmament of my thoughts and in my heart.

For Curie.

<b>INTRODUCTION.....</b>	<b>1</b>
<b>CHAPTER 1: EXAMINING EXISTING ACCOUNTS OF INTENTIONAL ACTION ...</b>	<b>6</b>
<b>Introduction.....</b>	<b>6</b>
<b>1. Control, reasons, and beliefs .....</b>	<b>8</b>
1.1. Rational explanations and action guidance in response to reasons.....	12
1.2. Knowing what you're doing and control .....	15
1.3. An example: Heuer's Belief-Control account.....	19
<b>2. Introducing the alternatives .....</b>	<b>21</b>
<b>3. Accounting for the fringes.....</b>	<b>25</b>
3.1. Automatic actions .....	26
3.2. Expressive Actions .....	31
3.3. Unskilled and/or "lucky" action .....	34
3.4. Summation .....	42
<b>4. Further difficulties for the existing alternatives .....</b>	<b>43</b>
4.1. Reason-backed: accounting for akrasia.....	43
4.2. Intention-backed: the problem of mutually exclusive intentions .....	46
4.3. Knowledge-backed: the problem of the lack of belief cases.....	49
<b>Concluding remarks.....</b>	<b>51</b>
<b>CHAPTER 2: MY CONTROL ACCOUNT OF INTENTIONAL ACTIVITY AND INTENTIONAL OUTCOMES.....</b>	<b>53</b>
<b>Introduction.....</b>	<b>53</b>
<b>1. System 2 Oversight.....</b>	<b>55</b>
1.1. Requirements of the intention-unique control mechanism .....	55
1.2. Dual Process Theory and System 2 Oversight .....	64
1.3. The belief to try.....	74
1.4. Intentional activity and the stream of behaviour.....	84
1.5. Accounting for intentional effects and consequences .....	93
1.6. Concluding remarks .....	101
<b>2. The unity of the three applications of the concept intention.....</b>	<b>102</b>
2.1. Introducing the three applications of the concept intention .....	103
2.2. Unity through control.....	106
2.3. Making sense of mutually exclusive intentions.....	113
2.4. Concluding Remarks .....	116
<b>CHAPTER 3: MORAL RESPONSIBILITY.....</b>	<b>118</b>
<b>Introduction.....</b>	<b>118</b>

<b>1.</b>	<b>Setting the stage: positioning my account .....</b>	<b>121</b>
1.1.	Merit versus consequentialist accounts.....	121
1.2.	Incompatibilism and compatibilism .....	123
<b>2.</b>	<b>Moral responsibility and intentional action .....</b>	<b>126</b>
<b>3.</b>	<b>The contested role of control: volitionist and attributivist accounts .....</b>	<b>132</b>
3.1.	Volitionism.....	133
3.2.	Attributivism.....	143
3.3.	Concluding remarks .....	150
<b>4.</b>	<b>Ambivalence and pluralism.....</b>	<b>151</b>
4.1.	Invariantism and Conservatism .....	152
4.2.	The problem of ambivalence cases and pluralist responses.....	157
4.3.	My dual-variant account of moral responsibility.....	163
4.3.1.	Attributability.....	164
4.3.2.	Accountability .....	167
4.4.	Concluding remarks .....	171
<b>5.</b>	<b>The nexus of control .....</b>	<b>172</b>
5.1.	Control condition on accountability: normal System 2 Oversight .....	172
5.1.1.	Normal functioning .....	174
5.1.2.	Accounting for consequences .....	181
5.2.	Concluding remarks .....	183
<b>6.</b>	<b>Consequences for the Doctrine of Double Effect.....</b>	<b>184</b>
	<b>Concluding remarks.....</b>	<b>190</b>
	<b>CONCLUSION .....</b>	<b>192</b>
	<b>BIBLIOGRAPHY.....</b>	<b>193</b>

## INTRODUCTION

In 1956, Elizabeth Anscombe famously protested the decision of an Oxford University committee to award an honorary degree to the then ex-president Harry Truman. Her reasoning, presented in a speech to her colleagues, was that by ordering the use of the atomic bombs on Hiroshima and Nagasaki, which resulted in the deaths of a great many civilians – some of them babies – Truman had committed a war crime. In light of this, she argued that one might as well “honor Genghis Khan, Nero, or Hitler” (Stoutland, 2011: 4). This was not a popular position at the time, and it remains a controversial one today. However, it is not Anscombe’s criticism of Truman to which I wish to draw attention, but rather one of the responses she received in his defence. This line of argument maintained that Truman was not morally responsible for the deaths of the civilians (or at least that this responsibility was of a much lower order), as he had not intended to kill them, and so their deaths were a regrettable, but unintentional, side-effect of deploying the atomic weapons. Along with this went the implicit assumption that the ex-president’s intention in ordering the use of the bombs, usually presented as the intention to end the war, was itself a morally permissible one, whose good consequences outweighed the bad. This kind of response confronts us with the intricate tangle that is the relationship between moral responsibility and the intentional status of human action, as well as the severity of the potential “real world” consequences of untangling this knot one way or the other. Indeed, lives can hang on this decision.<sup>1</sup> Determining the extent or presence of an agent’s moral and/or legal blameworthiness for his actions, is intimately bound to the question of what it was that the agent intended, and what it was that the agent did intentionally.

That there is *some* relationship between the intentional status of an action and an agent’s moral responsibility for it is about as uncontroversial as any claim about moral responsibility is likely to be. After all, as evidenced by the defence of Truman discussed above, our everyday practices surrounding moral responsibility are permeated with considerations of the intentional status of actions and consequences. Imagine the following situation: a man is reversing his car out of his driveway, and in

---

<sup>1</sup> One need only consider examples such as those of Oscar Pistorius (Mail&Guardian, 2013), Sanele May (SABC, 2013), and Gilberto Valle (The New York Times, 2013) to see this importance in action.

doing so he runs over a neighbour's child, killing her. It seems immediately apparent that he is causally responsible for the child's death. However, the question of whether he is morally responsible cannot be decided given only the facts of the case as presented thus far. It is my argument that the salient missing facts are those concerning the intentions of the driver, and the intentional status of his actions and their consequences. Consider now that we fill out this situation in one of two ways: (1) With Intention and (2) Without Intention. In (1), the driver, when it is demanded of him to explain his actions, states (truthfully), "I always hated that child, and so I intentionally reversed over her." In (2), when the same question is raised, he responds, "I had no idea she was there! I didn't do it intentionally!" It seems very clear that the agent in each of these versions of the situation are not on the same footing in terms of their openness to moral responsibility, and I will argue that it is the difference in the intentional status of the killing of the child in the two cases that explains this uneven footing.

At least partly in response to situations such as these, Anscombe embarked on her ground-breaking work on intention. For if the intentional status of the consequence of an action can play such a central role in determining the moral responsibility of an agent, then it is important to be clear about what we mean by such status, as well as when to ascribe it. Following Anscombe, I aim to investigate the relationship between moral responsibility and intentional action, at least as it pertains to individual responsibility. In other words, to do some work untangling this conceptual knot. The difficulty lies in the fact that disentangling the knot involves the consideration of a host of related notions, all of which stand in important relationships with intentional action, moral responsibility, or both. These notions – including but not limited to: control, agency, knowledge, moral reasons, free will, action, and desert – are not only complicated in and of themselves, but are also part of a web of mutual support and tension. This presents a challenge for any attempt at providing a simple account of any single knot within this web of knots, since any such account will, likely as not, implicate changes to several other significant relationships.

It is not my aim to map out the entirety of this web of interrelations. Rather the focus of this piece will be – to extend the metaphor – on only one part of the web, one knot in its network, one which I take to be crucial. I take this knot to be best expressed in

the form of the following question: is there a *necessary* relationship between an agent's moral responsibility for some action and the intentional status of said action. Developing an answer to this question requires the provision of convincing accounts of both the two central notions at stake: *intentional action* and *moral responsibility*. Presenting and defending my own accounts of each of these notions will form the primary tasks of my project, as I take it to be the case that once these accounts are in hand the relationship between the two notions is readily apparent. This dictates the structure of my dissertation: in Chapters 1 and 2 I develop an account of intentional action, while in Chapter 3 I develop an account of moral responsibility and demonstrate how these two accounts relate through the shared role played by control in both.

The first section of Chapter 1 examines the features that we can expect to find in control accounts of intentional action, and introduce what I label as the reason-backed, intention-backed, and knowledge-backed approaches to understanding intentional action – each named for what they take to be the characteristic feature of intentional action. These will then be tested by how they fare against the challenge of accounting for several difficult types of cases of intentional actions, what I call *fringe cases*. Although knowledge-backed accounts will be shown to fare best, none of the approaches will prove capable of providing satisfactory answers in all the cases. Finally, I tackle each approach in turn, raising what I take to be the most compelling objection to each.

In Chapter 2, I introduce my own account of intentional action. My account is what I call *control-backed*, as it takes the presence of a certain kind of control as the characteristic feature of intentional behaviour. It is also a *mechanism-focussed* account, in that I argue we should start our attempt to understand what is characteristic of intentional behaviour by investigating the actual mechanism through which an agent exerts her or his rational agency in behaviour. I begin with the identification of the requirements that must necessarily be met by such a mechanism. Having outlined these, *Dual Process Theory* is then introduced, a theory of human reasoning within the cognitive sciences in which our cognitive processes can be understood as falling into one of two systems (System 1 and System 2), each with its own distinct characteristics. I argue that the mechanism in question is best understood

as, and the previously discussed requirements best met by, System 2 oversight. This leads to my provisional formulation of the control condition on intentional action as *System 2 Oversight* (S2O).

I then move on to accounting for the epistemic condition on intentional action and widening the scope of my account beyond intentional action to encompass intentional omissions, and the intentional outcomes of actions and omissions. The first of these is achieved by defending the necessary role played by a belief to try in the proper formulation of S2O. The second requires the recognition that intentional actions are not in fact ontological particulars, but rather abstractions individuated from the stream of activity under the guidance of their intentional status. I will argue that the same is true for intentional omissions (which I will group with actions under the label of *activities*), but that intentional outcomes derive their intentional status somewhat differently, because of the difference in the directness of control exercised over activities and outcomes respectively. I also adduce support for my account of intentional activity and outcomes by showing that it can explain the unity of the three applications of the concept intention identified by Anscombe (1963).

Having completed one half of the two overarching tasks in the first two chapters, Chapter 3 has two central aims: To provide a convincing account of moral responsibility – or at least that part of moral responsibility that stands in a necessary relationship to intentional activities and outcomes, and to explicate the relationship between this account and that of intentional agency already developed. To achieve this first requires positioning my account relative to two live debates within the moral responsibility literature: that between merit-based versus consequentialist understandings of responsibility, and the longstanding argument about the relationship between free will, determinism (and indeterminism), and moral responsibility. I then clarify how my account treats the relationship between responsibility, praiseworthiness, and blameworthiness, and present a brief exploration of the complexities that emerge when considering the relationship between moral responsibility and intentional action, drawing out the salience of questions of control. The current discussion regarding the role of control in moral responsibility is then unpacked. This involves discussing the approaches of supporters and critics of a necessary role for control in moral responsibility, these being volitionists and

attributivists respectively. Having explored these existing accounts, I follow Watson (2004) and Shoemaker (2015) in pushing an objection applicable to accounts of both approaches: that no invariantist account of moral responsibility can capture the ambivalent nature of our responses to certain cases of moral responsibility. I also follow them in taking the best solution to this objection to be the endorsement of pluralism about moral responsibility. I introduce my own dual-variant account, that takes there to be two variants of moral responsibility: attributability and accountability. I further argue that only the latter of these has a necessary control condition, and that given the vital role of control in my account of intentional activities and outcomes, it is this variant of responsibility that stands in a necessary relationship with intentional behaviour.

Next comes unpacking the control condition on moral accountability, which I will argue is best understood as *normal* System 2 Oversight, a qualified version of the control condition on positive intentional status. I then extend this condition on accountability to intentional consequences. Due to the relationship between the control conditions on intentional agency and moral accountability, I argue that the relationship between moral responsibility and the intentional status of a behaviour is best understood in the following way: *an agent is open to being accountable for a given behaviour only if that behaviour is intentional*. Finally, I close out by returning to consider the motivating case, that of Anscombe and Truman. I identify that the defence of Truman is an example of the Doctrine of Double Effect, and argue that this defence is unsuccessful, as there is no morally relevant distinction as far as responsibility is concerned between an agent's intended outcomes and intentional outcomes solely based on whether they are intended or intentional.

## CHAPTER 1: EXAMINING EXISTING ACCOUNTS OF INTENTIONAL ACTION

### Introduction

To sufficiently untangle enough of the relationship between intentional activity and moral responsibility so that a necessary relationship between the two can be identified, it is, as I have mentioned in the Introduction, necessary to have a firm grasp of both central notions in the relationship. Sarah Paul makes a similar point when, in writing the entry on intention in *The International Encyclopedia of Ethics* (2013: 2658), she says the following:

The investigation of what we are morally permitted to do is integrally bound up with the puzzle of what it is to act. Intentional actions are paradigm objects of moral evaluation; therefore, grasping what it is to act is part of understanding and justifying such evaluation. In turn, the study of intentional action is integrally bound up with the notion of intention. What is done intentionally stands in some relation to the intention with which one acts: the very same physical event of an arm rising might on one hand be an unintentional spasm, and on the other any of the intentional actions of hailing a taxi, voting, stretching, or signalling for the revolution to begin. And in addition to contributing to the determination of what is done, the intention with which an action was performed may influence our moral assessment of that action. An account of the nature of intention and its relation to intentional action is thus highly relevant for moral philosophy.

The integral relationship between moral evaluation and intentional action that Paul has identified in this passage is what I called in my introductory chapter the conceptual knot of moral responsibility and intentional activity. She has also correctly identified that any account of intentional action will be inextricable from some account of the nature of intention.

Attempts to provide such accounts have been one of the philosophy of action's longest standing goals. The focus has been on attempting to ascertain the necessary

conditions of that sort of behaviour that is “the fruit and flower of the human will” (Ford 2011: 76). Such behaviour has – at least since Anscombe’s seminal discussion of the topic – usually been taken to be *intentional action*. For this reason, the attempt to develop an account of intentional action is well-trodden ground, as it is an enduring aim of contemporary action theory to provide a theoretically sound basis for distinguishing intentional action from the rest of human behaviour (Anscombe, 1963; Davidson, 2001; Mele and Moser, 1994; Seebaß, Schmitz, and Gollwitzer, 2013).<sup>2</sup> As should be expected, pursuing this aim immediately raises a host of questions: what is an intentional action? How should we understand the concept this term refers to? And how should we understand the relationship between intentional action and a host of related terms: intended, intention, intentionally, etc.? These are the questions anyone must answer if they are to present a convincing account of intentional action. An account, in other words, that is sufficiently grounded in our common language intuitions and usage to be relevant, as well as being philosophically rigorous enough to allow for the evaluation of these very intuitions and usages. This is a bit of a balancing act to be sure, and at some stages I will be favouring one side over the other. However, I will seek to show that there are good reasons to support my position in these cases, rather than it merely being a lack of balance on my part.

My aim in this chapter is to unpack and understand intentional agency by critically examining three influential approaches to making sense of intentional action. I categorized these approaches by what they take to be the characteristic feature of

---

<sup>2</sup> Depending on interpretation, this has plausibly been an aim of thinkers tackling questions about human action since at least Book III of Aristotle’s *Nicomachean Ethics*. Throughout the early parts of Book III, Aristotle sets out to “distinguish the voluntary and the involuntary”, with an eye to the role this distinction plays in questions of moral responsibility. Though the voluntary/involuntary pair do not align perfectly with intentional/unintentional, it is worth noting that there is not a clear corollary for the English term “intention” – in its modern usage – present in Ancient Greek. Indeed, the current use of intention is a comparatively recent development, possibly developing from Locke’s use of the term in Section 1, Chapter XIX, of *An Essay Concerning Human Understanding*, where, in the process of listing and describing modes of thinking, he states that: “our language has scarce a name for it: when the ideas that offer themselves (for, as I have observed in another place, whilst we are awake, there will always be a train of ideas succeeding one another in our minds) are taken notice of, and, as it were, registered in the memory, it is attention: when the mind with great earnestness, and of choice, fixes its view on any idea, considers it on all sides, and will not be called off by the ordinary solicitation of other ideas, it is that we call intention”. However, even this description is not quite a match for the contemporary use of the term, which is something of a partial amalgam of what Locke would have called “attention” and “intention”. This said, it is striking in reading Aristotle’s discussion in Book III how the dichotomy he sketches between voluntary and involuntary actions aligns enticingly with that commonly drawn between intentional and unintentional actions, particularly with regards to the role played by knowledge about one’s actions in determining voluntariness.

intentional action, i.e. that which chiefly determines an action as being performed intentionally. Each of these accounts proffers a different option for this central role, namely: reasons (understood as rationalizing explanations or judgements), intentions, and self-knowledge (usually understood as a special type of belief). I have labelled these three approaches *reason-backed*, *intention-backed*, and *knowledge-backed* accounts respectively. Importantly, I take all three these types of approaches to be examples of *control* accounts of intentional action insofar as they all hold that actions are only intentional if they are under some appropriate control, even if this is not intentional action's characteristic feature. As may be anticipated however, these accounts disagree on how this control is best understood.

I will begin my critical examination by first outlining some basic insights about intentional action and its relationship to reasons, beliefs, and the notion of control. This groundwork laid down, I then briefly introduce the three approaches to understanding intentional action. Having completed the expository part of the chapter, I then spend the rest of it in critique of the three approaches. First I see how they fare in accounting for a variety of fringe cases – instances of action where the intentional status might be difficult to discern, and where many accounts of intentional action give implausible answers. In particular, I will be looking at *automatic actions*, *Expressive actions*, and “*lucky*” *actions*. After reviewing the different approaches’ report cards, I then present individual criticism against each one: the problem of accounting for akrasia for reason-backed accounts, the problem of mutually exclusive intentions for intention-backed accounts, and the problem of cases where belief seems to be lacking for knowledge-backed accounts.

## 1. Control, reasons, and beliefs

The starting point for a control account of intentional action is the idea that what is characteristic and unique about intentional action is that it is action over which the agent has a certain kind of control or guidance. Let us call this certain kind of control *intention-unique control*.<sup>3</sup> This term serves as a theoretical placeholder, to be filled by

---

<sup>3</sup> This is not meant to imply that this kind of control is unique to only intentional action, there may be other things over which an agent could have this kind of control – such as judgements, perhaps, or certain second-order attitudes. However, intentional actions are uniquely those *actions* over which the agent has such control.

the correct conception of control once it is identified. The easiest way to see the initial appeal of this approach is to consider that the denial of control on the part of the agent usually serves to disqualify attributions of positive intentional status to her action: if I ask, “Why did you crash into that tree?” and your response is, “I lost control of the vehicle” it would be puzzling for me to conclude that you *intentionally* crashed into the tree. Similarly, though perhaps more controversially, if I ask, “Why did you steal that book?” and you (truthfully) reply, “I couldn’t control myself, I’m a kleptomaniac” it also seems that I would be in error to conclude that you *intentionally* stole the book. In these cases, the lack of control seems to be the result of some intervening factor preventing the agent from guiding her actions in the way necessary for intention-unique control. However, there are also cases of loss of control which results from an agent’s inability to guide their actions because of some epistemic shortfall: if I do not know that there is a vase on the table behind the door, and then open that door, striking the vase and causing it to fall, it would be peculiar to conclude that I *intentionally* knocked over the vase.

What is more, it seems reasonable to think that the presence of intention-unique control over some action is not merely necessary, but *sufficient* for it to be an intentional action. Showing this to be the case is much trickier than showing that intention-unique control is necessary for intentional action. After all, the provision of any number of cases where the agent has control and the relevant action is intentional does not yet secure sufficiency. However, it seems implausible to think of any case where an agent performs an action but declares, “though I had control over what I was doing, I didn’t do so intentionally.” Indeed, this looks a lot like a contradiction. If one attempts to flesh out what such a case may look like, it will probably be something such as the following: I was walking down the street and someone threw a ball at me, which I then, before realising what was happening, caught in my hand. The catching of the ball in this case seems to be within the agent’s control in at least some important sense. However, in any case like this we can question whether such actions are in fact intentional: consider if catching the ball leads to some bad consequence, say that the ball is wired such that upon being caught it detonates and takes off the agent’s hand. In this case, if the agent knew that the ball had this feature, and we asked the agent why she caught it, it would be perfectly intelligible for her to say, “I

didn't *intentionally* catch it, it happened before I realised what I was doing!" Is this not then a case of control without intentionality?

Against this possibility I will argue that though a certain type of control may be present in cases such as this, it is *not* the *kind of control* necessary for the action to be intentional. I explore what the right kind of control might be below. However, even if I am persuasive in my argument that the control present in such cases is the wrong kind of control, countering this single case can of course not establish sufficiency (nor indeed could any number of such cases), but I do present it as something of a challenge to those who think that there can be actions that are under an appropriate kind of intention-unique control without these actions being intentional, namely: to provide a case of action under such control but with absent intentionality.

The proponent of control accounts thus puts forward the following rough and provisional principle:

**Intentional Status Transmission (IST):** An agent's action is intentional iff the agent possessed the right kind of control over it.

The immediate follow-up question to this starting point is then: what is meant by "the right kind of control" here? Clearly not any kind of control will do. After all, even unintentional actions could be considered as being under an agent's *causal control*, in that the agent was causally necessary for the action to take place.<sup>4</sup> Consider, for example, a sleepwalker who is pouring in bathwater. Clearly some causal control is being exerted here, but at the same time the agent could, if made aware of her behaviour, compellingly claim that she had no intention-unique control over it, and that it was not intentional. This is what I argue is at work in a case such as the ball catching given earlier. When the catcher is asked why we should think of her action as unintentional, she is likely to respond with something along the lines of, "I couldn't stop myself!" This denial of control can take a variety of forms depending on the case: a denial of knowledge ("I didn't know the vase was there!"), a denial of intention ("I didn't intend to do that!"), or a denial of responsiveness to reasons ("It

---

<sup>4</sup> I will not be taking a stand on the causalist/anticausalist debate within this piece, as I take my account of intentional action to be neutral in this dispute.

didn't matter what I thought best, I was carried away by the wind!"). It might be thought that a denial of responsiveness to reasons must go hand in hand with a denial of intention, but this is not in fact the case. To see how this is possible, consider the following example based on one employed by Woolfolk, Doris, and Darley (2006): Two couples, Bill and Betty, and Max and Mary, are vacationing together at a resort. While there, Bill discovers that Betty and Max have been conducting an affair, and is greatly angered by this. On the flight back, the airplane is hijacked by a group of terrorists. These terrorists then experiment with a newly developed drug which, once applied, renders the target utterly physically obedient to the terrorist's commands. The target's thoughts and mind remain his or her own, but her or his body will respond in such a way as to obey the instructions of the terrorist. Bill is given the drug, and the terrorist orders him to kill Max. Now Bill in this situation knows that "he" is going to kill Max, and so cannot deny knowledge. We can also say, to make the point, that before the terrorists attacked the plane Bill had formed the intention to kill Max when the opportunity presented itself, and he has retained that intention. This means that he cannot deny having the intention to kill Max either. However, despite this, it would be wholly implausible to call his killing of Max intentional. What is missing is that Bill had no ability to modify his behaviour in response to reasons in this situation. This is the denial that makes it so that Bill's killing of Max is not intentional.

On the other side, intention-unique control cannot be something as stringent as conscious, reflective control – such as when an agent consciously deliberates about a course of action, decides upon it and then consciously implements it – as it is not difficult to think up cases of intentional action of which the agent is neither conscious nor reflective: such as absent-mindedly drumming on a table, or catching a cricket ball in the slips before the mind is able to consciously process having done so.

So, if this is what intention-unique control *isn't*, how should we understand what it *is*? To answer this, we must first take a moment to consider some crucial insights about intentional action. Particularly those pertaining to two important characteristics about intentional action that have been the most rigorously explored in the philosophical literature: (i) that intentional actions are characteristically related to acting for reasons, and (ii) that it involves some constitutive (possibly unique) epistemic

conditions. These clearly relate to two of the three forms of denial of positive intentional status that I mentioned above.

Much of the discussion of both characteristics has been given its contemporary foundations by Anscombe's (1963) discussion of intentional action. Famously, Anscombe said of intentional action that it is behaviour about which it is meaningful to "raise the question 'Why?' in the sense of inquiring into reasons for acting" (Bayne, 2010: 15). In other words, "what we do for reasons, we do intentionally" (Setiya, 2011: 171). Another way to state this is that an action is intentional when the agent, who performed the relevant action, when asked to explain why, provides a rational, rather than a causal explanation (Anscombe, 1963: 1-5). In addition, Anscombe (1963: 13-15) also claimed that intentional action is a subclass of those things that are "known without observation" – when they are known – by the agent, and that it is this special epistemic access to certain of our actions that is characteristic of them being intentional. The relevance of these insights for *intention-unique control* is that it indicates that such control will likely be best understood as some kind of ability to guide or direct actions in response to reasons, and that this control must either require or result in the agent having some noteworthy epistemic access to these actions.

### *1.1. Rational explanations and action guidance in response to reasons*

Beginning with the former insight first. I take the obvious first step to be Anscombe's observation that intentional actions are characteristically those actions for which the demand for a rational explanation has application. To understand the significance of this, it is first necessary to grasp the distinction between *rational explanations* and *causal explanations*. For her, the basis of this distinction begins with her consideration of a different distinction, that between *expressions of intention* and *predictions*. As pointed out by Driver (2011), there is some similarity between expressions of intention and predictions, insofar as both of them are "future-directed." Both seem to require a *belief* that a future state of affairs will occur. The difference Anscombe identifies is that, whereas we justify predictions through the provision of causal evidence, we justify expressions of intentions through the provision of reasons, a fact that appears to say a great deal about what makes intentional actions, intentional

(Anscombe, 1963: 1-5). Hence, for Anscombe, reasons and causes are distinct and not always interchangeable, at least when employed as means of justification. Still, the link between intention and prediction is a deep one. Indeed, it is arguable – and Bayne (2010: 5) does so argue – that intention requires prediction on a conceptual level. If the formation of intention requires a belief regarding the future, as Anscombe (1963: 5) certainly assumes it does, and prediction is the means through which we form beliefs about the future – or a belief we have regarding the future – then prediction is necessary for intention to have any content. I can scarcely form an intention if I am incapable of forming a belief concerning the future, since intention is always future-directed. Bayne (2010: 5) states this point more strongly when he says “[b]ut one thing seems to be certain: without, at least, a belief that we can make successful predictions, no intention will be formed.” This does not mean that intentions are derived in an unmediated fashion from our predictions (two people having the same predictions could very well form two disparate intentions), but that the ability to form intentions is at least partly founded on our belief in our ability to predict the future to a sufficient extent that our actions could change it in a meaningful way.

So, what is meant here by causal explanation and rational explanation? An illustrative example, provided by Driver, to clarify Anscombe’s meaning runs as follows:

[W]hen someone knocks a glass off of a table he may give an explanation that he saw a face in the window and that made him jump. This provides a causal explanation for why he knocked the glass off the table, but it doesn’t give a reason. The knocking of the glass off the table was not intentional, though it was caused by his being startled. (Driver, 2011)

A purely causal explanation of an action is one that provides no insight into the reasoning process of the agent causally responsible for it. This can be because no reasoning process took place, as in Driver’s example, or not. Even in cases where rational explanations are possible, causal explanations of a given behaviour can still be given. However, as Anscombe (1963: 10, 24) correctly stresses, what is explained through a rational explanation is not what is explained by a causal explanation, even if they are explaining the “same” behaviour. It is only rational explanations that reveal “something as *having a significance* that is dwelt on by the agent in his account, or as

a response surrounded with thoughts and questions” (Anscombe, 1963: 23). Rational explanations are explanations that employ or call for reasons – by which is meant here *normative reasons* – to explain actions. Whereas all behaviour for which there is a rational explanation can also be described with a causal explanation, not all behaviour that can be given a causal explanation can be given a rational one. Indeed, whereas all behaviour has a causal explanation, actions and omissions are the only types of behaviour that *can* have rational explanations, and it is precisely such actions and omissions that are intentional. It is also possible, in some cases, for rational explanations to serve as causal explanations, for example: my belief that I ought to tell the truth provides a reason for me to intend to tell the truth, *and could*<sup>5</sup> be part of the causal explanation of my action if I was to act according to my intention. However, the point remains clear: if an action (or behaviour more generally) *only* has a causal explanation, then it cannot be considered intentional. However, this does not mean that in every case where the Why? question has application there is a positive answer. In many cases the answer, “for no reason” is perfectly intelligible. It is not necessary that actual reasons can be provided in answer, what matters is that a question seeking such reasons has legitimate application. This will become very relevant when we turn to talk of the role of *capacity* in my own control account in Chapter 2: Section 1.1.

Given this relationship between reasons and intentional action, a control account can contend that a natural way to think about the kind of control we have over such actions is as the ability to guide or direct our actions in response to reasons. This would explain, in a straightforward manner, why intentional actions would always be open to the Anscombean “Why?” question, and usually have rational explanations. Anscombe herself did not make this argument, likely as she herself did not advance a reason-backed control account of intentional action, but the fact remains that such accounts can provide a very appealing story for why it is that rational explanations and reasons for acting are so intimately related to the intentional status of actions.

---

<sup>5</sup> Whether or not this will sound plausible will depend on your interpretation of the causal role of reasons in intentional actions, a point we can leave aside for this discussion.

### *1.2. Knowing what you're doing and control*

All accounts of intentional action make at least some assumptions about the epistemic conditions associated with such action – regardless of whether arguments for these assumptions are advanced – even if the assumption is simply that intentional action has no associated epistemic conditions. That said, it seems largely undisputed that performing an intentional action at least entails some epistemic demands on the part of the agent – even if these are not *conditions* on intentional action. Although the nature, quantity and quality of these demands are a matter of contestation, what is not generally contested (though I am not claiming it is *never* contested) is that an agent who performs an action X without any belief (conscious or unconscious) that she has done so cannot be said to have intentionally performed X, or to have X'd intentionally. Therefore, a question that a convincing account of intentional action must answer – be it a control account or not – is not *if* there are epistemic demands associated with intentional action, but rather: how much (and what) does the agent have to believe about herself (and the external world) for her to count as having acted intentionally (or intentionally omitted to act)? And does this knowledge have a special status? While pursuing an answer to this question, it should be kept in mind that most thinkers who have approached this topic have focussed not on the epistemic demands on intentional action, but rather on the epistemic demands on intention or intending. It is important to note that *having an intention* (or *intending*) and *acting intentionally* might have different associated epistemic demands.

It is also worth taking a moment to differentiate between what I have called here epistemic conditions, and what could be called epistemic requirements. Both are examples of epistemic demands associated with intending and intentional action, but they play different roles: the former, the epistemic conditions, must obtain in order for an agent to count as intending or acting intentionally. If these conditions do not obtain, then no positive intentional status is present. In contrast, what I am here calling epistemic requirements on intending and intentional action are normative rational requirements on intending and acting intentionally. Failure to meet one or more of these requirements does not entail the lack of an intention or of positive intentional status, but rather that the agent who fails in this way is guilty of criticisable irrationality as regards the intention or intentional action in question.

Attempts to provide accounts of the epistemic conditions on intending have tended to centre on discussing the relationship (or lack thereof) between *intention* and *beliefs*. There are three primary trains of thought: the first contends that to have an intention simply is to have a certain type of belief. This approach, famously taken by Harman in his 1976 work, “Practical Reasoning”, claims that intentions are a special species of beliefs.<sup>6</sup> This is then extended to intentional actions and consequences. In a similar vein, Setiya argues that the epistemic conditions for intentional action are not necessarily knowledge, but rather “justification of confidence, which comes by degrees” (Setiya, 2011: 174). An action can therefore be termed intentional if the agent had reasons for action and had the capacity to know what she was doing, even if this capacity was only realised in the form of a belief regarding what she was doing.<sup>7</sup>

The second train of thought holds that, though an intention cannot be reduced to a special kind of belief, such a belief forms a necessary component of intention. This was the kind of view espoused by Davidson (2001) – who held intentions to be a complex of a belief and a desire – and it was the dominant position in the conversation for some time.

The final train of thought holds that intentions cannot be reduced to beliefs or to belief-desire pairs, since it takes intentions to be irreducible mental states on equal footing with beliefs and desires, and does not assume any necessary relationship between having any particular belief and having an intention. Consider that one of the signature qualities of certain types of pathological behaviour is that the apparent intentions upon which an agent acts, and/or some aspects of the reasoning processes that lead her to a given apparent intention, are not known to the agent. This raises the possibility that an agent might have an intention that they do not believe themselves

---

<sup>6</sup> Examples of thinkers who defend this approach include Velleman (1989) and Setiya (2007; 2010). Velleman does not in fact call intentions beliefs, but “self-fulfilling expectations that are motivated by a desire for their fulfillment and that represent themselves as such” (1989: 109), however it is not at all clear to me what an expectation might be other than a belief that something will come about. When I have an expectation that the waiter will bring my food, this seems to be a belief. When I have an expectation that I will successfully vault a fence, this seems like a belief. As such, and since I do not think that anything I argue for hinges on this, I will be grouping Velleman together with Harman and Setiya as the representative proponents of the view that intentions are beliefs.

<sup>7</sup> Setiya (2010: 174) presents this argument as part of his novel defense of what he terms “Anscombe’s Principle,” which claims that “If A has the capacity to act for reasons, she has the capacity to know what she is doing without observation or inference – in that her knowledge does not rest on sufficient prior evidence.”

to have (see Bratman, 2009a: 30). However, beliefs still play a significant role in accounts that follow this assumption, but due to their involvement with the rational requirements on intending (which I discuss below), rather than as conditions for the presence of intending.<sup>8</sup>

Separate from, but not unrelated to, how one goes about answering the question of the epistemic conditions on intentional activity, it is widely agreed that there are certain rational requirements regarding the consistency of intentions to beliefs. Requirements that place an intending agent under some rational normativity, and that open the agent up to charges of criticisable irrationality if they are not adhered to. The simplest example of this is the requirement that for an agent to hold an intention rationally, it is necessary that the agent must not believe that achieving it is impossible, however, it is acceptable if the agent has no particular belief on the matter, one way or the other. In a similar vein, Bratman (2009b: 413) states the requirement as follows: “*Intention Consistency*: The following is always pro tanto irrational: intending A and intending B, while believing that A and B are not copossible.”<sup>9</sup> Another important requirement is that of *instrumental rationality* (also called *Means-Ends Coherence*), which holds that it is irrational to hold an intention A, believe that B is the means to A, and then not either forsake the intention to A or the belief that B is a necessary means.

However, there is strong disagreement as to how exactly these requirements should be understood, and how exactly they relate to intention. Whereas Bratman and Holton, for example, argue that the rational requirements on intention are unique to intention, or at least cannot be explained by rational requirements on other mental states (such as beliefs), Velleman and Setiya disagree. They argue that at least some of the rational

---

<sup>8</sup> Bratman (1987; 2009a; 2009b; 2013) and Holton (2009) have both put forward influential recent examples of accounts of this type.

<sup>9</sup> This is not the only formulation of Intention Consistency that Bratman has employed. In earlier work he presented the requirement as follows: “it should be possible for my entire plan [or intention] to be successfully executed given my beliefs are true. This is the demand that my plan [or intention] be *strongly consistent relative to my beliefs*” (Bratman, 1987: 31). Both formulations are meant to convey that any intention must endeavour to be consistent with the agent’s beliefs. To intend what I believe is impossible is irrational, and part of what makes a given mental entity an intention is that it should strive to avoid such irrationality. Of course, this does not prevent me from having irrational intentions, but it indicates that any intending agent must strive to meet this criterion, even if she fails. This *striving* is an essential entailment of intention. Although Bratman claims that I do not have to have a belief to A in order to have an intention to A, this still means I must “check” my beliefs whenever an intention is formed. Or, more accurately, I would run the risk of being justifiably criticisable as irrational if I did not.

requirements on intention can be explained by the rational requirements governing our beliefs – i.e. the requirements of theoretical reasoning – since on their accounts for an agent to have an intention to X is simply for that agent to have a special kind of belief that she will X (in the case of Velleman, 1989) or for her to have increased confidence that she will X (in the case of Setiya, 2007). This view, which has been dubbed *cognitivism about intention*, has become the centre of much of the discussion concerning intention’s relationship to practical and theoretical reasoning. Cognitivism can come in various forms: a strong account may argue that intention is a species of belief and endorse cognitivism about the demands of practical reason, arguing that they can be reduced to those of theoretical reason. A more moderate account might be one where intention is understood as a type of belief but cognitivism about practical reason is not adopted. And finally, weak cognitivism could be extended to include any account of intention that sees intention as entailing a belief (Levy, 2017: 1). For my purposes in this dissertation, I will be using the term “cognitivism” to refer to the position that intentions are beliefs. Any convincing account of intention – and given their intimate relationship, any convincing account of intentional action – must at the end of the day make a ruling on whether cognitivism is the appropriate way of understanding intention.

In principle, a control account of intentional action is reconcilable with all three of the primary trains of thought, as well as either cognitivism or non-cognitivism about the requirements on intention. Whether intentions are beliefs, belief-desire pairs (or other composites containing beliefs as a part), or unique and irreducible mental states, each one makes possible the central role of control. If intentions are beliefs, or belief-desire pairs, then the control in question concerns how responsive our actions are to these beliefs and desires. In other words, we appropriately control the action when our intention (belief, belief-desire pair) guides our conduct appropriately. Michael Smith (1994, 2011, 2012) is a good example of a proponent of this variety of account. He argues that action is separated out from the rest of behaviour by the guiding role played by desire-belief complexes. On his account desires are taken to dispose an agent to a certain action, with beliefs as to how to accomplish the content of a given desire fulfilling a guiding role. He considers this combination to be the “foundation of agential control” (2011: 81). Interestingly, since Smith takes these desire-belief

complexes to also be the originating source of an agent's reasons to act, Anscombe's insight about the applicability of the Why? question is arguably captured here.

Though this is fairly straightforward if intentions are taken to be belief-desire complexes, there are justifiable question marks about how an intention understood as a pure belief could guide conduct, given that beliefs are usually not assumed to have motivational power. Much of the argumentation surrounding cognitivism in fact rides on this point. Velleman (1989) – as already noted, a cognitivist – argues that intentions can guide conduct despite being only a belief because these beliefs are self-referential, and we are motivated to make our beliefs of this sort true. So, the intention to X is understood as being a belief with a content something like “I am going to X in virtue of this very intention to X”, and we are motivated to X through a pre-existing motivation to make this belief true. In opposition to this, those who hold that intention is a unique mental state, discrete from desires or beliefs, have argued that intentions are intrinsically motivating, and are by their very nature conative states. Such views will tend to promote understanding intention-unique control as pertaining to the guidance of our actions by our intentions. We will return to this argument later, but for now it is worth noting that the conflict exists, and that a control account could fit with any of these views but will have to take a side.

### *1.3. An example: Heuer's Belief-Control account*

To see how this all comes together in a control account, I will briefly outline an example of one, namely that of Ulrike Heuer. I choose her account both because it is a recent example, and because I take it to be a very compelling articulation of a knowledge-backed control account.

In her account, Heuer takes the type of control required to be *belief-control* (2014: 297). For an agent to exercise this type of control over her actions it is necessary that she have a self-referential belief regarding the action taking place: that is, the agent must have the belief that she is in fact doing what she is doing. This belief need not be consciously held, but should be tokened in the agent's belief box. Furthermore, the control that the agent exerts over the action in question must not merely coincide with this self-referential belief, but must follow “in virtue of” the belief. So, an action I

perform would be intentional if the control that I had over the action obtained “*in virtue of*” my having the relevant self-referential belief (Heuer, 2014: 298). So, say that I am typing a sentence. Heuer points out that this behaviour could be either an unintentional action, or an intentional action.<sup>10</sup> According to the belief-control account, this would be an unintentional action even if I was in control of my typing the sentence and I believed that I was doing so, but my belief played no role in my control of the action. To see her point, consider breathing. My breathing may be under my control, and I may believe at a given moment that I am breathing, but, says Heuer, my breathing would still not be intentional.

Though Heuer never makes exactly clear what this “*in virtue of*” is meant to mean, it does become clear that she takes one of the virtues of her approach to be that it explains the role of reasons in intentional action. Her contention is that in order for an agent to be responsive to a reason for her to act, or not to act, she must have the belief that the reason in question is indeed a reason for her to act or not. In her (2014: 301) own words: “In order to respond to a reason (that a person believes she has), she must know what she is doing (or at least have a belief about it).” In other words, belief-control is a necessary condition for reasons-responsiveness. Heuer takes this to explain the intentional status of cases where there may be no reason why the agent is performing the action, but they are intentional insofar as the agent *could* respond to reasons to stop them. So, imagine a doodler who becomes aware that her doodling is disrupting her class. If we grant that she takes this as a reason for her to stop, then, provided she has belief-control, she should be able to do so. On the other hand, a similar doodler who lacked belief-control would be unresponsive to reasons to stop, and so would not have control. This allows us to arrive at her control condition:

**Belief-control condition:** an agent’s action X is under belief-control – and so intentional – iff (i) she has a self-referential belief that she is performing X, and

---

<sup>10</sup> It could also conceivably be mere behaviour, or a “mere event” (Ford 2011: 76). The latter would be the case if the movements of my hands and fingers were being caused by the direct interference of another agent (say someone has tied me to the chair and is moving my hands and fingers to type the sentence), while the former would be the case if the movements are the result of some pathological process, as we may say of the compulsive movements of a person with severe Tourette’s syndrome. However, these possibilities do not impact the thrust of Heuer’s argument.

- (ii) she is sufficiently responsive to the reasons for and against X *by virtue of* having this belief.<sup>11</sup>

Having presented this brief summary of Heuer's belief-control account, I must stress that I take her account to be an important step in the right direction. In her paper, she mentions that Bratman "has pointed out to me, I am in danger of mixing two 'traditions' of understanding intentional agency: some have argued that knowledge or belief is crucial to intentional agency; others suggest that control or guidance by intention is", adding that "[b]ut I hope that I have shown that we need to go beyond guidance by intention to understand the unity of intentional agency. It may thus turn out that the two traditions are not incompatible after all" (2014: 301). I am in full agreement with the possibility she raises here – that the two traditions can indeed be shown to be compatible if unified under a control account, however I take her own approach to fall short of this. Whereas her account puts forward an understanding of control that is centered on the role played by self-referential belief in allowing for reasons-responsiveness, my own account reverses this relationship: it is a fact of how our reasons-responsive mechanisms function that they necessarily entail certain beliefs whenever they culminate in intentional actions. However, before discussing my own account in more detail in Chapter 2, I will first consider the dominant alternative control accounts present in the literature – including Heuer's – and explain why I take them to fall short of being convincing.

## 2. Introducing the alternatives

Control accounts of intentional action primarily come in three influential flavours, three ways of understanding the control necessary for intentional action. These are what I will be calling *reasons-backed*, *knowledge-backed*, and *intention-backed* accounts. To be convincing, a control account of intentional action should not only outline its own approach, but also explain why it is to be preferred over these alternative views. With this in mind, I will briefly discuss each of these in turn, before presenting my arguments against them.

---

<sup>11</sup> It is worth noting that Heuer is unsure whether or not this condition will also work for non-human animals. She does think that non-human animals can act intentionally, but is unsure if they can be said to act for normative reasons. She sets the worry aside by noting that in the animal case a different explanation of belief-control might be called for.

However, before proceeding it is important to note that the thinkers who articulate the views that I am labelling as reasons-backed, etc., do not usually use the label “control” when describing their accounts, yet at the same time each of them talks in their own way of “agential control” (Smith, 2011, 2012), or “guidance” (Bratman, 1984), or “motivated in the right way” (Velleman, 1989), or “caused in the right way” (Davidson, 2001). Heuer is an exception to this trend, speaking about control directly, which was in part the reason for my use of her account as an example. All these listed notions I take to be gesturing toward the role of control. While this may be self-explanatory in the case of, “agential control” and “guidance”, it may be less so with the latter two. The key lies in the important phrase *in the right way*. The answer to which kind of actions are in fact motivated or caused in the right way line up with the answer to the question “which kinds of actions does the agent have appropriate control over?” To see this, consider why for Davidson, for example, it is important to stress this notion of the *right kind of causation*, namely: the problem of deviance. We can imagine a case where an agent has the appropriate desire-belief complex, and that this complex causes the relevant action, but that this causation is of the wrong sort. In a widely-employed example, consider a man who is intending to signal his confidant by shaking, but the fact that he has this intention causes him to become nervous and start shaking, which his confidant takes to be the signal. There have been many attempts made to resolve the problem of deviance, but these are not important to the discussion here, what is important is that the easiest way to explicate why it would seem like a problem to label the shaking in this case as intentional, is that the agent did not have control over it, at least not intention-unique control.<sup>12</sup> Consider one influential attempt at resolving this problem put forward by Setiya (2003: 348):

[T]he crucial concept is that of *guidance*. When an agent  $\Phi$ 's intentionally, he wants to  $\Phi$  and this desire not only causes but *continues to guide* behavior towards its object. [my emphasis]

---

<sup>12</sup> Similar points can be made with talk of *motivating in the right way*.

Given this, I take the explanatory lacuna gestured to by phrases like “in the right way” and “in virtue of” in these contexts to be best filled with some conception of control.<sup>13</sup>

Reasons-backed accounts, the most influential of which is almost undoubtedly that of Donald Davidson (2001), who argues that intentional action is explained by the fact that all intentional actions are actions that are explained in the appropriate way by the agent’s reasons. Indeed, he argues that intentions are themselves a kind of rational judgement (2001: 39). Reason-backed accounts (roughly) argue that what makes a given action intentional is that the action is performed for reasons, that is: an agent’s action is intentional if it is the case that the agent’s “behaviour can be predicted and explained through the attribution to them of beliefs, desires, and *rational choice* [my emphasis]” (Millican and Wooldridge, 2014: 4). Though this choice need not be understood as a consciously deliberative one.

In contrast, intention-backed accounts, of which Michael Bratman (1984; 1987; 2013) is a noteworthy proponent, posit a necessary relationship between any instance of intentional action and some intention. Intention-backed accounts contend that for a given action to be intentional, it must stand in some relationship to an intention. For reasons I will mention below, this relationship needed not be direct – as in, to intentionally X an agent must have an intention to X – but it is the case that this account argues that there cannot be cases of intentional action without the agent having some intention.

Knowledge-backed accounts, which are often cognitivist accounts, usually contend that intentions are a certain kind of belief, though what sort of belief varies between

---

<sup>13</sup> It may be thought that anticausalist accounts of intentional action will avoid these problems entirely by denying that intentions are causes of intentional action, but are rather constitutive of them. However, though they do sidestep the issue of causal deviance, the problem of potential deviance is not in fact wholly avoided. On an anticausalist account it will usually be said that an agent intentionally X’s if they X *in virtue of* reasons. So, my slipping and falling to the ground is intentional only if I slipped and fell to the ground for some reason I took myself to have – perhaps I was performing in a play for example. Yet, I could take myself to have decisive reason to perform some action, and then that action could take place, but yet we would not describe the action as intentional. For example: I am driving toward my brother’s house with the intention of killing him. I take myself to have a decisive reason to do so. While driving I see a man in the road, yet I am so set on killing my brother that I simply run the man down, killing him. Unbeknownst to me though, the man in question was my brother. It would be implausible to say that I intentionally killed my brother here (I intentionally killed a man, certainly). For this reason, it is important that the action – to be intentional – must be done in virtue of the reason in question, not simply be co-existent with the reason, or even follow from it in the wrong way. I take the presence of control to fill this explanatory space as well.

the different approaches. What is central to these accounts, however, is that actions are deemed intentional if it is the case that the agent has the right sort of belief or knowledge about the action. For many who follow this approach, their starting position is Anscombe's previously mentioned insight that a distinguishing feature of intentional action is that the agent performing said action has non-observational knowledge of what they are doing, if they have knowledge of what they are doing at all.

These brief introductions are not intended to be treated as full explanations of these different positions, but rather to relay their basic commitments and tenets. Instead, the fuller scope of the arguments advanced by each of these approaches will be explored throughout the next section, where I test how the different accounts fare in making sense of difficult cases. It is my thinking that this method will more organically introduce not only the details of these accounts, but also how they seek to defend themselves against potential counterexamples and limitations.

These accounts should not be thought of as entirely incompatible with each other. Consider Heuer's belief-control, which clearly gives a necessary role to both beliefs and reasons. Though her account does not take intention to be a belief, she in fact makes no claims on the nature of *intention*, her self-reflective belief could be understood as something akin to that present in Velleman's account of intention. And though she does not require that every intentional action must be backed by a reason, she does require a sensitivity to salient reasons. So how do we adjudicate what type of account is under discussion? By asking three questions and considering the answers: (I) Is it the case on the given account that every instance of intentional action demands an intention, a reason, or a belief. In the case of Heuer's account, the answer is clearly a belief. It is for this reason that I take Heuer's account to be a knowledge-backed control account. She takes the presence of a “kind of control [that] makes it possible to modify one’s action in the light of reasons” (2013: 201) to be what is characteristically necessary for intentional action, and from this starting position concludes that *only* control that involves a self-referential belief can fulfill this role. All instances of belief-control involve the presence of a self-referential belief, but, important to note for later, not all instances of the presence of a self-referential belief involve belief-control. (II) If more than one of these are necessary for intentional

action, is there some priority involved? For example, on Anscombe's account of intentional action, though the role of reasons is important and she seems to hold that at least openness to reasons for acting is necessary for intentional action, she gives priority to an agent's privileged self-knowledge of what you are doing as the key characteristic of intentional action. For this reason, I think it appropriate to take her account to be an example of a knowledge-backed account of intentional action. One clear way in which such priority can be discerned, is if one of these features follows from another. If X is necessary for intentional action, and X always results in Y, then Y will also necessarily be present in all cases of intentional action. Y may even play its own necessary role, rather than being only correlated with the intentional action. Even so, in this case I will take feature X to have pro tanto priority over feature Y. Or (III) if one of the features is necessary and sufficient, rather than merely necessary, while the other feature is only necessary, then I will treat the former feature as central.

This sets the table for us: to support a particular control account of intentional action against its competitors, it is necessary to either show that what they take to be the characteristic feature of intentional action is not in fact necessary, or that it is subordinate to some other feature, or that it is merely necessary whilst some other feature is necessary and sufficient.

### **3. Accounting for the fringes**

Most accounts of intentional action (Anscombe, 1963; Davidson, 2001; Velleman, 1989; Harman, 1997; Setiya, 2007; Bratman, 1984, 2013) – indeed I have not found an exception – attempt to first account for clear and paradigmatic cases of intentional action, and then after this has been accomplished, set out to make sense of cases at the fringes. So, whereas all the primary contenders can give a fairly satisfactory explanation for why my typing this sentence is intentional, they differ on cases such as aimless and absentminded doodling. Of course, part of the difficulty in accounting for the cases at the periphery is that it is not always clear if these should be considered cases of intentional action at all. I mentioned in the introduction that there is a balancing act to be undertaken between what reconciles best with the facts of how we employ these terms, and the aim of theoretical consistency. Though the former must be considered carefully and not disregarded lest we lose sight of the very phenomenon

we are trying to explain, it is also the case that common usage can be in error, or in need of correction. For the purposes of my arguments then, I will take it as a theoretic virtue of an account if it captures more of the common usages than its competitors, and wherever a usage is excluded (i.e. our labelling of a given instance of action as intentional action is taken to be an error in our use of the term) a compelling reason should be provided for it. In what follows I will test how the three primary alternative control accounts fare in accounting for the fringes.

The three types of fringe cases I will be discussing are: (1) cases of automatic action (which can come by degrees), (2) cases of Expressive Action, and (3) cases of unskilled or lucky action.

### *3.1. Automatic actions*

Cases of automaticity in actions are usually directly contrasted with actions undertaken under conscious or deliberative control (Schlosser, 2013; Moors and De Houwer, 2007; Norman and Shallice, 1986). These actions are said to be *automatic*, and an agent can often be unconscious of the details of the performance of the action, the underlying mechanism or process resulting in the action, or (more rarely) of the entire action itself. There are many flavours of automatic action, Schlosser (2013: 215) gives “over-learned motor skills, automatic stereotype activation, and automatic imitation” as examples, to which can be added actions such as absentmindedly drumming one’s fingers, or so-called “slips of action” or *action slips* (Norman and Shallice, 1986: 13). Along with this variety of type, automaticity in action is also thought to come by degrees. By this is meant that conscious or deliberative control is not an all or nothing feature of an action. This is quite vividly exemplified in the process of over-learning a motor skill: when I first begin playing tennis I exert a lot of conscious control over my shots, as I improve and practice, I cede more and more of the control to unconscious processes. Eventually, if I over-learn the motor skill sufficiently, these processes may even kick in without any conscious control on my part at all.

On the face of it, automaticity is worrying for all the competing accounts of intentional action: if the process guiding the action are unconscious, does it make

sense to think of them as responsive to reasons? It is normally thought that *agents* are responsive to reasons, not sub-agential processes – particularly unconscious ones. There is also the problem about degrees. Automaticity comes by degrees, but the presence of reasons backing an action does not. Though we can speak of actions that have more or fewer reasons in favour of it, this does not coincide with degrees of intentionality at all. An agent can act clearly and unequivocally intentionally, and do so for only one reason. Alternatively, it can be noted that reasons can come in different degrees of strength, with some reasons favouring an action more strongly than others. Yet this again does not coincide with the degrees of automaticity: as I continue to over-learn some skill, X, the strength of my reasons to X need not change at all, yet over time the degree of automaticity involved may increase, and my degree of control diminish. On a straightforward reason-backed view then, an action is either backed by a reason and so is intentional, or is not backed by a reason and is not intentional, there cannot be degrees of intentionality as required in the automaticity examples.

If an action is sufficiently automatic, then it also seems as if no intention is likely to back the action. Certainly, when I automatically switch my grip on my tennis ratchet prior to making a shot, it would seem odd to say that I have an intention to do so. However, intention-backed accounts can respond to this by following Bratman in jettisoning the “assumption of tight fit” (1984: 394) concerning the relationship between intention and intentional action. He describes this assumption as follows:

They both [the Simple View and the Volitional Thesis<sup>14</sup>] assume that if there is a distinctive pro-attitude involved in intentionally A-ing, it will be a pro-attitude specifically in favor of A – that there must be a tight fit between what is done intentionally and what is intended (willed).

By jettisoning this assumption, it becomes possible for an agent to intentionally X while not requiring an intention to X. In his own words:

---

<sup>14</sup> The Simple View is Bratman’s (1984: 377) coverall term for accounts of intention that endorse the assumption of tight fit. The Volitional Thesis is another attempt at providing a convincing account of intention that Bratman criticises in “Two Faces of Intention” (1984). I will not be examining the Volitional Thesis in my arguments here, as I take Bratman’s criticisms of this kind of account of intention to be accurate, and as such I take the Volitional Thesis to have already been shown to be unconvincing.

I propose to give up the assumption of tight fit and to distinguish between what is intended, and the sorts of intentional activity in which an intention may issue.

*Making this distinction, we can say that when I A intentionally I intend something, but I may not specifically intend to A* [my emphasis]. Our notion of intentional action embodies a complex scheme for the classification of actions (or, perhaps, actions ‘under a description’). To understand the relation between intention and intentional action we must recognize that the factors that determine what is intended do not completely coincide with the factors that, on this scheme, determine what is done intentionally. (Bratman, 1984: 394)

This is not to say that if an agent has some intention then any action they take is necessarily intentional. It is not enough to have the temporal conjunction of intention and action. Bratman’s idea is that there will be some set of criteria that specify which actions would count as intentional given the presence of a certain intention, though he never spells out what exactly he takes this set of criteria to be. To understand how this works, consider a case where an agent is playing cricket and standing in the slips. She has the intention of catching the ball if it comes by her, which will involve stretching to the left or to the right. Of course, if the ball goes left and she stretches right she will fail to catch it, and vice versa. The ball is nicked by the batsmen and flies past her left side, at this moment stretching to the left is a means to fulfil her intention, and if she was to stretch left and catch the ball this would be an intentional action. Yet, at no point need she have formed the intention to stretch left – indeed given the speed of cricket balls off the bat, there may not have been sufficient time to do so.<sup>15</sup>

Another way of understanding this relation is to consider Bratman’s (1984: 395) notion of the “motivational potential” of an intention. He defines this notion as follows:

The notion of motivational potential is intended to mark the fact that my intention to B may issue in my intentionally A-ing, not to explain it [the account of how this issuing occurs]. It is a *theoretical placeholder*: it allows us to retain theoretical room for a more complex account of the relation between intention

---

<sup>15</sup> See McCann (2005) for an example of a proponent of an intention-backed approach who defends the ability of the Simple View to account for problematic cases such as these.

and intentional action while leaving unsettled the details of such an account. Such an account would not itself use the notion of motivational potential but would, rather, replace it with detailed specifications of various sufficient conditions for intentional conduct. (Bratman, 1984: 369)

Bratman goes on to provide a deeper explanation of how he envisages these “various sufficient conditions” to work in his 1987 *Intention, Plans, and Practical Reason*. Finkelstein (2005: 588) provides a good summary of the solution offered up in this book as follows: “Only those effects that lie within the “motivational potential” of one’s action should be thought of as done intentionally, meaning that a person must have *consciously adverted to* and *actually deliberated on* [my emphasis] an effect for it to count as something done intentionally.”

The crucial thing is that an agent can X intentionally, even if she doesn’t have an intention to X, provided she has an intention to Y, and X meets the afore-mentioned set of criteria. I will return to both the idea of jettisoning the assumption of tight fit and the notion of motivational potential in Chapter 2: Section 2, as it has important consequences for understanding the unity of the three guises of the concept intention. However, for our purposes here, the benefit of taking a wide fit view is immediately apparent: I may not have had an intention to change my grip, but I certainly had an intention to make the shot, and so my changing my grip can be intentional, provided the relationship between making the shot and changing my grip meets the criteria mentioned above. Bratman’s (1987: 121) preferred example is that of a jogger training for a marathon, who realises that in doing so he will be wearing down his shoes, and decides to run anyway. In this case, he argues, the running down of the shoes is intentional, and this can be made more apparent if we think of the shoes as having some sentimental value to the jogger.

However, there are still cases of automatic action that are problematic for even wide fit intention-backed accounts. These are cases of action undertaken automatically that are normally thought to be intentional, but seem to involve no intentions whatsoever. Heuer (2014: 264-265) provides an example of a case like this which she employs as a counterexample to reasons- and intention- (though not knowledge-)backed accounts:

*Doodling*: Finally, certain things we do when passing the time are intentional, but not done for reasons. Doodling while listening to a philosophy paper is intentional behaviour, but we don't normally doodle for a reason... There is no intention to doodle, and presumably no other intention either — certainly no intention to try.

Consider also a case such as an agent who, while walking down the street, thoughtlessly kicks a stone down the road. In this case as well, there is seemingly no intention at all.

Knowledge-backed accounts are somewhat shielded from most of these problems, as such accounts usually claim that the belief necessary for intentional action need not be a consciously held one, and need merely be accessible to the agent under certain conditions – such as on reflection, or to meet the demands of theoretical reasoning. This accessibility can also come by degrees, which allows such accounts to make sense of the degrees of automaticity. This makes sense of cases such as over-learned skills, as well as doodling and stone-kicking. However, on this view, behaviour during action slips would not count as intentional. An action slip is a situation where an agent sets out with an intention to perform some task X, but falls into undertaking some alternative action Y without conscious awareness of having done so. In most cases the action Y will be some habitual action often undertaken. To illustrate this, consider the example of a man who plans to go to the store, and as part of doing so intends to go to his shed in order to pick up his shopping bags. However, this man is a fervent gardener who frequently goes to his shed to don his gardening gloves, and when he arrives in the shed he falls into this routine and begins donning the gloves. When the man realizes what he is doing, he is startled and surprised, he had not realized what he was doing.<sup>16</sup> What is noteworthy about this case is that the agent did not possess a belief about what he was doing that could be reflexively recalled – he *did not have knowledge of what he was doing*.

However, it is not at all clear whether an agent's behaviour in these action slips *should* be considered intentional. It is plausible that cases such as these should simply

---

<sup>16</sup> I owe this example to Berys Gaut, though he would likely disagree with my verdict as to the intentional status of such action slip behaviours.

not be deemed to be intentional actions. This certainly seems like it would have been Anscombe's response. After all, any knowledge that an agent might gain of these action slips would be achieved through observation. Consider for example a case where the action slip results in some harm being caused: a barber is very used to delivering a certain cut for a certain client, to the point where this has become habitual. Yet one day this client enters the shop and asks for a different cut. Unfortunately, in this case, the barber falls into an action slip, and delivers the regular cut. In the face of the client's angry accusations – let us imagine that his new haircut threatens his chances in a job interview – it is not at all implausible that the barber might honestly declare, “I didn't do it intentionally!” Thus, it appears that knowledge-backed accounts can provide at least sufficiently convincing responses to automatic cases.

### 3.2. Expressive Actions

An action is an Expressive Action when the agent in question expresses emotion through the action, but not in order to do so. Hursthouse (1991: 58) provides a number of examples where an agent expresses some emotion through an action: such as throwing an “uncooperative” tin-opener on the ground or kicking doors that refuse to shut or cars that refuse to start. Of course, it is often the case that agents express their emotions through far more thought out and considered action: a dancer might practice for weeks or months in order to perfect a routine that, when properly executed, is a means through which he expresses some emotion. Or the writing of a poem, or buying a loved one some flowers. What is meant to make Expressive Actions different is that the action not only expresses emotion, but that this expression is unreflectively direct, and that the action is not undertaken *in order to* express the emotion. It is not a metaphorical or planned expression of the emotion, but a visceral and *emotional response*.

As with the automaticity cases, it seems that the reasons-backed approach is once again on the ropes. It is implausible to say in all these cases that the agent acted how she did for a reason, given that “reason” is understood as a normative reason. It may, of course, be the case that the individual's actions are in line with normative reasons that are present in the area, but this is not the same as *acting for* a reason. Consider: it

may be the case that kicking the non-starting car may somehow (as seen in so many films) cause the engine to start. If the agent kicked the car *in order to* bring this about, then this would not be an example of Expressive Action as understood here.

An intention-backed account is in a better position, since at least some, if not all, Expressive Actions may well meet a wide fit account's criteria for positive intentional status. In other words, it may be the case that the agent has the intention to throw the tin-opener on the floor, even if she does not have an intention to thereby express any emotion. Provided that the action in question is appropriately related to this intention, then a wide fit account could explain its intentionality.

Heuer, though conceding these cases, argues that there remain some instances of Expressive Action that are not amenable to this treatment. The example she puts forward is that of “banging the table in frustration” (2014: 265). It seems that she must contend that an agent cannot bang the table in frustration while having an intention to strike the table but no intention to express emotion by doing so. It is puzzling to think why this must be the case. What makes the instance of table banging different to that of, say, throwing the tin-opener? In both cases the emotion in question could be understood as frustration, and it does not seem more implausible for an agent to have an intention to strike the table than it is for her to have the intention to throw the tin-opener. Perhaps Heuer thinks the point of difference lies in the fact that the throwing of the tin-opener seems to be more goal-directed than the striking of the table. Yet, I would argue that this appearance is deceiving. Striking the table is no more or less an aim than throwing the tin-opener. A defender of Bratman's account could easily point out that both agents are more than likely to answer yes to the question, “did you *intend* to X?” It is usually thought that to intend to X means to have an intention to X.<sup>17</sup> Though much does depend on the criteria for determining which actions count as intentional given the agent's possession of a given intention, it seems at least plausible that intention-backed accounts can explain the positive intentional status of Expressive Actions such as these.

---

<sup>17</sup> To see the persuasiveness of this, consider how absurd it would be to say of an agent that she *intends* to kick the ball, but has no *intention* to do so. Or that she has an *intention* to pick the flower, but she does not *intend* to pick it.

Knowledge-backed accounts are again less concerned by these fringe cases. In all the Hursthouse examples, as well as those put forward by Heuer, it is unproblematic to think that the agent in question has the belief that she is in fact doing X. The tin-opener thrower, the table banger, the car kicker, etc., all likely believe that they are undertaking the given action. At least insofar, it would seem, as these actions can be called intentional. Indeed, an advocate of such an account could compellingly argue that it is precisely the presence of the correct type of belief that allows us to discriminate between those Expressive Actions that are intentional and those that are not. Consider, the advocate might say, that there are surely some cases of Expressive Action that we would not want to call intentional: the curling of a lip in disgust, staring angrily, clenching a fist in rage. In these cases, it seems that at least sometimes the action is not intentional. But how to tell them apart? One possible answer would be that in the cases that are intentional the agent has the belief that she is undertaking the action, even if this belief is not consciously held (but is sufficiently accessible). This may be best illustrated by comparing the following two vignettes:

*No Belief*: Matthew is a strongly politically active Democrat in the United States, and is watching the results of the 2016 Presidential Election as the final tallies in Michigan are coming in. As he realises that his preferred candidate has lost, he is filled with anger, and unthinkingly clenches his fist. His wife, Miriam, enters the room and asks him why he has clenched his fist so tightly. Upon hearing the question, he looks down and realises he clenched his hand. He replies, “I didn’t realise I was doing it, I am just so angry about the result, I guess my rage got the better of me.”

*Belief*: The case unfolds in the same manner up till the point where Matthew hears the results. In this scenario, he again clenches his fist, and believes he is doing so, even if he does not consciously entertain the belief at the moment. Miriam again enters and asks her question. This time Matthew does not have to look down to realise that he was clenching his hand, he knows that he was. His answer would likely be, “because the news has just made me so mad!”

While Matthew (*Belief*) seems to have clenched his fist intentionally, it would be implausible to say the same of Matthew (*No Belief*). This is at least pro tanto evidence

in favour of the knowledge-backed approach to understanding intention-unique control.

### 3.3. Unskilled and/or “lucky” action

Unskilled and/or lucky actions are those where an agent brings about an outcome, but does so despite a very low probability of success. Examples include: entering and winning a lottery, randomly pressing numbers on a keypad and striking the correct combination to open a door, or a terrible marksman taking aim at a difficult to strike target and successfully doing so. Though not all such cases are ones where the agent in question is unskilled, we are generally inclined to call these actions “lucky”, though not in the sense of fortunate: it could be the case that an agent bring about some end in a “lucky” way but that this end turns out to be a bad one (for the agent in question or others). The trouble with these actions – the reason they are often located at the periphery of intentional actions – stems from the strong intuition that if a given outcome being brought about is sufficiently lucky, then this undermines its positive intentional status, or from the fact that many of the features usual of intentional action seems to be present, yet it seems implausible that the action in question is intentional

The examples for this final type of peripheral case that I will mostly be employing are drawn from the work of Mele and Moser (1994), though they deployed these examples to a different effect than I will be.<sup>18</sup> It appears to me that this variety of

---

<sup>18</sup> The aim of their use of the examples was to serve as evidence for their particular account of intentional action, which they give as the following (Mele and Moser, 1994: 63):

Necessarily, an agent, S, *intentionally* performs an action, A, at a time, t, if and only if:

- (i) at t, S A-s and her A-ing is an action;
- (ii) at t, S suitably follows – hence, is suitably guided by – an intention-embedded plan, P, of hers in A-ing;
- (iii) (a) at the time of S's actual involvement in A-ing at t, the process indicated with significantly preponderant probability by S's on balance evidence at t as being at least partly constitutive of her A-ing at t does not diverge significantly from the process that is in fact constitutive of her A-ing at t; or (b) S's A-ing at t manifests a suitably reliable skill of S's in A-ing in the way S A-s at t; and
- (iv) the route to A-ing that S follows in executing her action plan, P, at t is, under S's current circumstances, a suitably predictively reliable means of S's A-ing at t, and the predictive reliability of that means depends appropriately on S's having suitably reliable control over whether, given that she acts with A-ing as a goal, she succeeds in A-ing at t.

fringe cases can be broken down into three rough categories: (I) those where the agent has *correct beliefs* about the means to their end, but success is *statistically unlikely*, (II) where the agent has *incorrect beliefs* about the means to achieve their end but through attempting what they take to be the correct means they in fact perform the actual means and the end is achieved, and (III) the agent has *incomplete beliefs* about the means to achieve their end, or *inadequate ability* to meet these ends reliably, but through attempting what they take to be the correct means they in fact perform the actual means and the end is achieved.

Each of these can be associated with an example: (I) Brandon is a basketball player. In a particular match, he finds himself in a position to take a very difficult shot, with a very low likelihood of success. Brandon knows that his odds of making the shot are faint, but takes it anyway. In fact, he makes the shot. (II) Aaron has been diligently trained at Assassin School to believe that he can only kill his target by shooting him through the heart, and believes this to be so. During an assassination, Aaron fires on his target, intending to strike him in the heart and kill him, but misses and strikes him in the head, killing him. (III) Nick works at a nuclear power plant. One day, some malfunction threatens to result in a radiation leak from the main reactor. Nick knows that to avert the crisis, he needs to input a ten-digit code on a keypad on the first try. Unfortunately, Nick did not have clearance to know the code, and so takes a guess and inputs ten numbers hoping to strike on the correct combination. As it turns out, Nick gets lucky and inputs the correct code, preventing the leak.

Another way of considering these cases, is that they represent situations of *partial control*, where the reasons for the control being partial differs in some details, but always involves some gap in the relationship between the agent and the means they are undertaking toward an end. For Brandon the lack of control follows directly from the unlikelihood of success, or more fundamentally, from his lack of control over the chances of success. The more capable an agent is of insuring the success of an action, other things being equal, the greater the agent's control over it. If we say, for instance, that if Brandon were to practice hard for ten years then he would be capable of

---

I take (i) to be clearly correct, (ii) and (iv) to be on the right track – though (ii) is problematic in that it demands that there be some related intention in play – and (iii) to be unnecessary once the role of intention-unique control is properly accounted for.

making the shot in question easily, then we would also say that after this practice he now has greater control over the outcome of the action. Aaron seems to be lacking control because his striking the target in the head represented a failure to achieve what he incorrectly believed to be a necessary subordinate goal toward killing the target: hitting him in the heart. Finally, Nick lacks control because he has less than complete knowledge of how to insure his action's success. The overarching theme is that the more the agents must rely on the co-operation of variables outside their control for the success of their actions – either because of the difficulty of the action itself or a lack of sufficient knowledge – the less such success seems intentional.

For reason-backed accounts the difficulty is not that in these cases there are no clear reasons present, but rather that in several instances it would seem that the actions undertaken are undertaken for a reason, but that they should not be considered intentional. In the case of Brandon, it is certainly possible to provide a rational explanation of his making the shot. And taken at face value it is easy to have the intuition that he made the shot intentional. However, if we modify the case by stressing the low chance of success, say by stipulating that the chances of Brandon making the shot was 1-in-1 million,<sup>19</sup> then this intuition begins to seem less clear. And that is the point: the degree of unlikelihood seems to have an influence on whether we tend to think of the Brandon's making the shot as intentional, even if all other facts remain unchanged. One of these unchanged facts is that the action is backed by a reason, which is a problem for the reason-backed account.

When it comes to Aaron the assassin, it is plausible to say that killing the target does not have a rational explanation since the target's death was an effect of Aaron

---

<sup>19</sup> Consider a comparison with a lottery case. If Brandon had entered a lottery that he had a 1-in-1 million chance of winning, and then indeed won, there is something strange about saying that Brandon *intentionally* won the lottery. As Sliwa (2017: 129) notes: "Winning the lottery (as opposed to scamming it) just isn't the kind of thing that one can do intentionally." This said, there is an important disanalogy between Brandon's case and that of a lottery, namely: in Brandon's case, he has a greater opportunity to influence the outcome through his relevant skills – that is to say, he has a greater opportunity to exert control over the outcome – than in the case of the lottery. Even if Brandon's skills are only mediocre, this will still improve his odds of success, whereas in the example of the lottery, there is presumably no such parallel. Does this mean that Brandon's making the shot is intentional? I would argue no, since the influence of his control is simply insufficient in degree, but it is noteworthy here already – and will be discussed more in Chapter 2: Section 1.5. – that in the lottery case no amount of practice, training, or knowledge (provided that the lottery outcome is indeed random) can improve the agent's odds of success, whereas this is not the case for Brandon.

shooting him in the head, and he did not shoot the target in the head for a reason – indeed he had “good” reasons, or at least reasons he took to be good, to avoid shooting the target in the head. If this is the case, then reasons-backed accounts should argue that killing the target was not intentional. This seems to be a plausible response, so the reason-backed account seems to handle this case. It is important, as Mele and Moser stress, that we not be influenced by the moral valence of Aaron’s behaviour. Even if we conclude that Aaron did not intentionally kill the target, this does not mean that he is necessarily morally off the hook for having done so.

In the case of Nick, it is plausible to say that he did not intentionally avert the reactor leak, though it certainly seems correct to say that he both intentionally *tried* to avert the reactor leak and was intentionally inputting the ten-digit code. What seems in doubt is whether he intentionally input *the correct code*, since he did not know what the correct code was. So, can Nick’s inputting of the correct code be given a rational explanation? This is a difficult question to answer. On the one hand, Nick certainly input the correct code for a reason. But on the other, there does not seem to be any reason why he would have chosen the exact set of ten numbers he in fact did. It seems that the reason-backed account can give us a plausible answer here: that Nick did not intentionally input the correct code, but that he did intentionally input a ten-digit code that proved to be correct.

Overall then, reason-backed accounts can do a decent job of tracking our everyday attributive practices concerning these cases when the lucky action involves inadequate beliefs (either false or incomplete ones), but provides an implausible answer in the case where control is diminished due to the sheer difficulty or unlikelihood of the outcome.

Turning to the intention-backed account, there is a similar difficulty in dealing with the case of Brandon. After all, Brandon surely has the intention of making the shot, and this intention certainly guides his attempt in making it, as unlikely as his success might be. Yet we are inclined to say that Brandon, at least in the million-to-one case, did not make the shot intentionally, it was simply too lucky for that. This may be a problem for the intention-backed approach as it raises the possibility that there can be cases of action that follow in the appropriate fashion (there is no deviance problem in

this case) from an agent's intention(s), but still be unintentional. However, wide-fit intention-backed accounts do have a potential solution. In discussing motivational potential, Bratman (1987: 121) provides us with the following rough example of sufficient conditions to be met in order for an agent's behaviour to fall within an intention's motivational potential:

S intentionally A's if:

- (1) S wants to A and for that reason intends to try to A; and
- (2) S A's in the course of executing his intention to try to A; and
- (3) S A's in the way he was trying to A; and
- (4) conditions (2) and (3) depend, in an appropriate way, on S's relevant skills

If we plug in the variables from Brandon's case, it can be argued that his making the shot does not meet condition (4). But much hinges on how "in an appropriate way" is to be understood. After all, making the shot did depend to some degree on Brandon's skill. Were he an infant the shot would have been impossible. As I discussed at the beginning of Section 2, these appeals to "the appropriate way" or "the right way" are usually best understood in terms of control. It should also be noted that failures of condition (3) will also often represent failures of control. It seems then that the central question is whether or not Brandon had *sufficient control* over the outcome. His skills, beliefs, and desires all play a potential role in determining this. Setting these intricacies aside for the moment, it is at least reasonable to think that Bratman could give a plausible explanation of Brandon's case.

Turning to Aaron, the immediate question to ask is, "what intention is Aaron really acting on?" Is it the intention to kill the target or the intention to shoot him in the heart, or both? It seems uncontroversial to say that Aaron does hold both intentions, and that Aaron intends to shoot the target in the heart *as a means to* kill him. The intention to kill the target (in addition to Aaron's admittedly mistaken beliefs) explains why Aaron intends to shoot the target in the heart. But in fact, Aaron struck the target in the head, and he had no intention to do this. But did killing the target by shooting him in the head fall within the motivational potential of Aaron's intention to kill the target? If we consult the list of criteria Bratman provides, Aaron's killing his target meets condition (1) and (2), but does not meet conditions (3) and (4). And the

failure to meet (4) follows from the failure to meet (3). Aaron did not kill his target *in the way he was trying to*. Given this, and given that (3) seems to be a reasonable criterion, an intention-backed account can give a compelling answer here.

How intention-backed accounts respond to Nick's case is similar, but importantly different to Aaron's case. Again, the agent can be thought of as having an intention to avert the leak, and has the subordinate intention of trying to input the correct code. Using the list of criteria Bratman provides, Nick's inputting of the correct code meets all the conditions listed. Conditions (1) and (2) follow straightforwardly, and it does seem correct (though less clearly so) to say that Nick input the correct code *in the way he was trying to do so*. On this account then it looks like Nick *did* intentionally input the correct code. This is less compelling than in Aaron's case, and seems an implausible answer. Though it may at a stretch seem correct to say that Nick intentionally averted the disaster, there is something decidedly misguided in saying that he intentionally input the correct code.

Knowledge-backed accounts usually argue that in order for an action to be intentional the agent must have a particular kind of belief pertaining to it. If the given proponent of such a position is a cognitivist – and so takes having an intention to purely be having a certain kind of belief – then in many respects her account would mirror that of the proponent of an intention-backed account. For instance: there is again the idea that in order for an action to be intentional a certain kind of mental state must be present, and that the action must be linked to this mental state in some appropriate manner. And in a somewhat analogical way to how the move to a wide-fit intention-backed view allows for the recognition of actions as intentional that are not necessarily backed by an intention to perform that exact action, knowledge-backed proponents such as Setiya (2008) have argued that what is needed for an action to be intentional is not always a belief aimed at the action in question. It is sufficient if the agent is performing the given action *by* performing some other action for which the condition does holds. In his own words (*ibid.*: 319):

If A is doing  $\phi$  intentionally, A believes that he is doing it or is more confident of this than he would otherwise be, or else he is doing  $\phi$  by doing other things for which that condition holds

Let us call this set of conditions *Knowledge (Belief)*. A further similarity is that, when faced with cases like that of Aaron's and Nick's, Setiya appeals to a notion of things having to be done in the “right way” and according to the agent's plan (2003: 363). This is clearly reminiscent of Bratman's condition that the action must be brought about in the way that the agent planned to bring it about, and again introduces a central, but somewhat vague, notion of control. Setiya's argument is that if this control is not present, then the reasons to act that led to the relevant intention-as-belief does not transfer to the action in question, and so the latter cannot be deemed intentional. Given these similarities, it should then probably be unsurprising that knowledge-backed accounts understood in this way would provide similar responses to lucky cases as intention-backed ones, at least in the cases where the agent has false or incomplete beliefs. Because of this last point, I will be working in a different order than usual and will first be discussing Aaron's and Nick's cases, before turning to Brandon's.

Looking at Aaron's case, we can see that following Setiya's conditions above we would have to conclude that his killing the target was not intentional, as it did not “run through [his] conception of *how* [he] will bring about the end” (2003: 363). In the case of Nick, the answer depends on what Nick's actual beliefs are. If Nick believes that he will succeed in inputting the correct number (which would be very irrational given his lack of any evidence for this belief) then his inputting the code would be straightforwardly intentional. If Nick does not believe that he will succeed in inputting the correct code, then his action in doing so can only be intentional if he inputs the code by doing something that he does believe he is doing. An obvious option might seem to be an intention to try to input the correct code, *a la* Bratman's method, but Setiya prefers to avoid the use of “trying” as he claims this is “*not* enough... [the agent] is and must be *doing* [my emphasis] specific things” (*ibid.*: 343). However, the result is the same, as what we can say is that Nick was inputting digits, and believed that he was doing so, and that by doing so he input the correct code. Following from this, Nick should count as having input the correct code intentionally. This is a strikingly counterintuitive result, as it was for the intention-backed account.

Looking at the case of Brandon, there are again two possibilities: first, the situation where, as the ball is passing through the air, Brandon believes that he will make the

shot – in which case his doing so counts as intentional directly, as he is doing  $\varphi$  and believes that he is doing so. And, second, the situation where Brandon does not believe that he will make the shot. In this latter possibility, Brandon’s action is still intentional since he will presumably have the appropriate belief about moving his arms and pushing with his fingers to satisfy *Knowledge (belief)*. This, again, is a very counterintuitive answer.

Partly in response to possibilities such as these, and partly because of his commitment to the view that agents have non-observational knowledge of their intentional actions, not only merely beliefs (which can be unjustified), more recently Setiya has made a move away from *Knowledge (Belief)* toward a different set of conditions, ones that do not concern an agent’s beliefs, but rather their (sometimes non-propositional) *knowledge-how*. Adjudicating the possibility of non-propositional knowledge-how is outside the scope of this work, and as assuming its possibility only serves to strengthen one of my competitor’s accounts, I will be making this assumption from here on. According to Setiya (2012: 287), *Knowledge (Belief)* is best replaced with the following, call it *Knowledge (How)*:

If A is doing  $\varphi$  intentionally, then A knows how to  $\varphi$ , or else he is doing it by doing other things that he knows how to do

To see if *Knowledge (How)* represents an improvement, let’s have a look at the cases again. The verdict in Aaron’s case remains unchanged, though in this case it’s because his killing his target did not properly follow from his knowledge of how to do so, rather than his beliefs only. However, things change up in the case of Nick, and become muddy in the case of Brandon. To understand Setiya’s view on Nick’s case, consider his response to a similar case: imagine that you’re trying to defuse a bomb, and you’ve narrowed down your options to a handful of wires. Cutting the single correct wire will disarm the bomb, cutting any incorrect wire will set it off. Running out of time you make a call and cut a certain wire. Fortunately, you guessed correctly and the bomb is disarmed. In this case Setiya argues that you defused the bomb intentionally, due to *Knowledge (How)*, but that you did *not* cut the correct wire intentionally as you did not know that was what you were doing when you did it. Though I disagree with this verdict – I will argue that both Nick and the bomb-

disarmer do not intentionally fulfil their aims – it is certainly an improved answer to this case, lending credence to the idea that Setiya may be right in his shift to knowledge-how.

If we look at Brandon's case again, the question now hinges on whether or not it would be accurate to say that Brandon knew how to make the shot. Setiya provides a direct answer when he writes (2012: 296): “we can equate knowledge how to  $\varphi$  with being disposed to  $\varphi$  when one so intends. This disposition may depend on propositional knowledge of means.” It seems that Brandon's disposition to make the shot will depend on his chances of succeeding, and so we have a spectrum where there is presumably a threshold of likelihood below which a disposition is no longer present – indeed, Brandon would then have a disposition to fail – and so his action would not be intentional. Though this is a less jarring answer than the one that came before, there is a concern here that the scope of intentional actions might be unacceptably narrowed. It seems wrong, for example, that if Brandon's odds were 3-to-1 against his making the shot that if we were to make it, it would not be intentional, yet in this case Brandon would still be disposed to fail. This is not a reason to forsake the knowledge-how enterprise, but it is a reason to develop a better notion of control in order to make sense of cases like Brandon's.

### *3.4. Summation*

Having completed our tour of the periphery, let us consider how our three accounts have fared. The reason-backed approach has the poorest showing, struggling to account for automatic actions and Expressive Actions, though faring better with lucky actions. Though all these accounts agree that reasons play an important role in understanding intentional action, it does not seem plausible to think that the actual guidance of a reason to act is what is characteristic of intentional action. Not only are there intentional actions that do not seem to have such reasons, but there are also cases of agent's X'ing for a reason which are not examples of intentionally X'ing. Intention-backed accounts fared somewhat better, being able to account for Expressive Actions and a good deal of automatic ones – though there remains a difficulty with actions that seem to have no related intention at all. In terms of lucky actions, the intention-backed approach, when enhanced with Bratman's refinements,

can give a good explanation for cases where an agent brings about an intended outcome in a manner other than those they attempted, a plausible answer in the case of pure unlikelihood, and a counterintuitive answer in the case of incomplete beliefs. Lastly, knowledge-backed accounts give a strong showing of themselves, handling automatic and Expressive actions without much trouble. Lucky actions proved more challenging, but after considering Setiya's updated knowledge-how account, only Nick and Brandon's cases proved elusive – and even then, I would argue that in the former it provides an improved answer and in the latter its shortcomings result primarily from a flawed notion of control. It certainly appears that knowledge-backed accounts are positioned to be the biggest rival to my own. Turning away from the fringes of intentional agency, in the next section I explore what I take to be the most compelling general criticisms applicable to each of the alternatives.

#### **4. Further difficulties for the existing alternatives**

##### *4.1. Reason-backed: accounting for akrasia*

As was highlighted in Section 1.1., it is commonly accepted that guidance by normative reasons for acting is an important part of understanding intentional action, regardless of whether it is viewed as intentional action's characteristic feature or not. After all, intentional action seems to reflect something about what an agent takes to be a reason for action, which in turn reveals elements of the agent's values, reasoning process, and attitudes. Taking the further step to a reason-backed account of intentional action is therefore not a surprising move. Davidson provides what is undoubtedly the most influential of such accounts. The crucial move in his enterprise is to identify intentions with *all-things-considered judgements*.

These are rational judgements about what is desirable, and are based on an agent's attitudes and beliefs. For Davidson, they constitute an agent's primary reason for acting. This action need not in fact eventuate, but the agent's actions must be explicable in terms of this judgement. In other words, an agent who claims to have an intention to X, but whose behaviour does not indicate that they in fact have an unconditional judgement toward X, does not in fact have an intention to X. On his view, for an action to be intentional it must be such that it is explained in the proper

way by such a rational judgement. Davidson argued that such a judgement emerges from an agent's attitudes and beliefs, even if these are not consciously grasped at the time. As he says, “[the agent] must have attitudes and beliefs from which, had he been aware of them and had the time, he *could* have reasoned that his action was desirable” (2001: 85). For Davidson, I must be aware of my all-things-considered judgement – i.e. my intention – although not necessarily of what has led me to hold that rational judgement.<sup>20</sup>

However, this understanding of intention as an all-things-considered judgement faces serious problems when confronting instances of akrasia. After all, if I must be aware of my judgement, but act against it, then my behaviour would not be appropriately explained by the judgement, and would fail to count as intentional. Yet this is an absurd outcome, given the existence of intentional akratic action. Recognising this shortcoming, Davidson argues that all-things-considered judgements, must be thought of as *unconditional judgements*.

To grasp how Davidson thinks about this kind of judgement, consider the following:

Every judgement is made in the light of all the reasons in this sense, that it is made in the presence of, and is conditioned by, that totality. But this does not mean that every judgement is reasonable, or thought to be so by the agent, on the basis of those reasons, nor that the judgement was reached from that basis by a process of reasoning. There is no paradox in supposing a person sometimes holds that all that he believes and values supports a certain course of action, when at the same time those same beliefs and values cause him to reject that

---

<sup>20</sup> The reason that Davidson takes intention to be an all-things-considered judgement is because anything less than this opens the way toward potential contradictions. Consider the following example he discusses (2001: 98-99): I have a pro-attitude toward sweet foods, and this leads to a *prima facie* judgement that eating sweet foods is desirable. If we take such judgements as being sufficient to be an intention, then it will also be a reason for so acting. I also have a pro-attitude to avoid poisonous food, which gives rise to an analogous judgement and reason for acting if we assume that a *prima facie* judgement can count as an intention. However, this opens the way for a contradiction, since there are some foods that have both characteristics. This means that an agent would end up holding that a given course of action is both desirable and undesirable. Given this, Davidson (2001: 98) contends that:

It is a reason for acting that the action is believed to have some desirable characteristic, but the fact that the action is performed represents a further judgement that the desirable characteristic was enough to act on—that other considerations did not outweigh it. The judgement that corresponds to, or perhaps is identical with, the action cannot, therefore, be a *prima facie* judgement; it must be an all-out or unconditional judgement which, if we were to express it in words, would have a form like ‘This action is desirable’.

course of action. If  $r$  is someone's reason for holding that  $p$ , then his holding that  $r$  must be, I think, a cause of his holding that  $p$ ...[But] his holding that  $r$  may cause his holding that  $p$  without  $r$  being his reason; indeed, the agent may even think that  $r$  is a reason to reject  $p$ . (Davidson, 2001: 40-41)

What he is getting at here is that, although an all-things-considered judgement is a judgement where “all truths, moral and otherwise” are taken into account, this actually means only “the sum of all that seems relevant to him [the subject]” (Davidson, 2001: 40). This would mean that even if all the reasons I take to be relevant indicate a certain choice, I could still choose to act otherwise based on considerations (reasons or motivations) that I have not, and importantly could not have, *consciously* considered. These unconsidered considerations are opaque to me. In a rather delightful conclusion Davidson remarks on this: “What is special in incontinence [*akrasia*] is that the actor cannot understand himself: he recognizes, in his own intentional behaviour, something essentially surd” (*ibid.*: 42).

Although this defence succeeds in countering the original criticism, it opens Davidson's position to a different problem. This problem is that, if Davidson is correct, an agent can never hold an intention opposed to the balance of reasons, taken to include unconscious reasons. This means that an agent can never perform an intentional action that is irrational by her own reasoning, if we take her unconscious reasons into account. For example: After consideration of all the reasons I hold relevant I take A to be a better choice than B, but I still choose to intend B rather than A. According to Davidson this must be because I have an unconditional judgement that B is in fact better than A. This means that I may, at the time, think that I am acting irrationally in the sense that I intend B even though I have no reasons (that are transparent to me) for doing so. However, if the unconscious reasons were explicated, then this would no longer be a case of me acting irrationally, as in fact I did – and had to – act in accordance with the balance of reasons that includes my unconscious reasons.

The issue is that Davidson cannot allow us to fail to intend in line with our rational judgement, as this would require that intention be something other than just such a judgement. This is not convincing, however, as it is entirely plausible that I can act in

opposition to the balance of reasons (which is exactly what *akrasia* is taken to show), be this balance transparent to me or not. A related issue is that Davidson's account problematically widens the scope of intentional action: the behaviour of a kleptomaniac in grabbing an item off the shelf, an agent's behaviour during action slips, an agent flinching away from an oncoming strike despite intending not to flinch, all of these would count as intentional under Davidson's account. And yet in each of these cases the agent in question could, seemingly justifiably, declare, "I didn't do it intentionally! It was involuntary!"

Reason for the lack of persuasiveness of this outcome of Davidson's account can be found in a consideration of Anscombe's original insight concerning the applicability of the "Why?" question, particularly the fact that this demand for a rational explanation was intimately bound to the fact that such explanations revealed something about the agent's reasoning process, it revealed "something as *having a significance* that is dwelt on by the agent in his account, or as a response surrounded with thoughts and questions" (Anscombe, 1963: 23). It is unclear how considerations that are entirely epistemically opaque to an agent can play this crucial role. Such considerations might very well be causes of action, but in order to serve the rationalising and revelatory role described by Anscombe at least some degree of epistemic accessibility is needed.

#### *4.2. Intention-backed: the problem of mutually exclusive intentions*

The most significant criticism of intention-backed accounts – apart from the possibility of cases of intentional action that have no related intention even in the wide sense – is the problem of mutually exclusive intentions. The best way to understand the problem is probably still the gamer example provided by Bratman in his 1984 work, "Two Faces of Intention". He uses the example of a gamer playing two games simultaneously, in each case trying to strike a target (T1 and T2 respectively) with a missile. The games are set up so that striking either of the targets makes it impossible to hit the other, so although the gamer is trying to hit both the targets, he cannot hit both. The problem here is that it seems that the gamer cannot intend to hit T1 *and* intend to hit T2, as these are two mutually exclusive goals. In no way does approaching T1 bring me closer to T2 or vice versa. The question then

becomes, how can we then make sense of the common intuition that as he is playing, the gamer intends to hit T1 and intends to hit T2, and that if he successfully strikes one then he did so intentionally, and – crucially – that in undertaking this course he is not being irrational?

The problem that the gamer example highlights is that, while we would normally consider cases of holding mutually exclusive goals irrational, it is unlikely that we would call the gamer in the example irrational. If the gamer sets up the games as Bratman describes, and then attempts to hit T1 and T2, we would not, in everyday life, take such behaviour to be an example of criticisable irrationality. Yet, on an intention-backed account, any given intentional actions must be related to a relevant intention. This immediately raises the spectre of looming contradiction. In order to retain the intuition that the gamer is not irrational, while still embracing an intention-backed account of intentional action, Bratman unsurprisingly argues that we should jettison the assumption of tight fit, which was discussed in Section 2.1. As a brief reminder, the assumption of tight fit held that in order for an agent to intentionally A, it was necessary that the agent had the intention to A. In place of this assumption Bratman rather argues that we should consider every intention to have a motivational potential – a useful theoretical placeholder for the set of criteria that should determine whether a given action is deemed intentional or not. This means that some action A could be deemed intentional not only if the agent in question had the intention to A, but also if the agent had the intention to B, and A fell within this intention's motivational potential.

However, this does not entirely resolve the problem. It seems that my goal of hitting T1 or of hitting T2 is not only something I might do intentionally, but is also a *source of rational requirements*. Most clearly, the requirements of instrumental rationality. My aim of hitting T1 pressures me to select the correct means toward that end, and just so for the aim of hitting T2. These pressures are usually associated with intentions. It then seems reasonable to think that, in the course of trying to bring about the overarching intention of striking one of the targets, the agent also develops subordinate intentions to strike each one individually, as these are necessary stepping stones that play a crucial role in guiding my intentional actions. But this again leads to the possibility of an agent holding an irrational combination of intentions.

To try to avoid this possibility, Bratman finds it necessary to introduce talk of “settled objectives” (Bratman, 2009a: 18-19).<sup>21</sup> He does this to resolve what he sees as a continuing threat of rational contradiction in cases like the gamer example. On this view, in cases where I have a plan that seemingly entails two intentions that are mutually exclusive, such as “I intend to hit T1 and T2” where I cannot achieve both, each individual element should be considered to be a “settled objective” rather than an intention (Bratman, 2009a: 19). So, I can say that my plan entails the general intention-for-the-future “I intend to hit at least one target,” or even the disjunctive, “I intend to hit T1 or T2,” but this overall intention entails that I have a settled objective to hit T1 and a settled objective to hit T2.

These settled objectives still engage the instrumental rationality requirement (or some similar requirement), but not the requirement of consistency that would apply to an intention. Since settled objectives do not have the same rational requirements that intentions do, the fact that these objectives are mutually exclusive is not a concern. Also, any subplans I might form to achieve these objectives would still count as intentional. For example: “I intend to press the fire button to hit target T1.” By distinguishing settled objectives from overall intention in this way, Bratman hopes to avoid the problem of rational contradiction. I contend that this is unnecessary, as well as both opposed to our linguistic intuitions and theoretically undesirable. I am reluctant to accept this line of reasoning precisely because in our everyday talk I would not refer to my “settled objective” to hit T1 or T2; rather, I would refer to my *intention* to hit T1 or T2. Furthermore, it hinges on the creation of a placeholder concept, *settled objective*, that seems to have no content of its own, engages precisely those rational requirements that are desired and none of those that are problematic, and is introduced purely to make a perceived contradiction go away. At the very least this solution should be avoided if there are other possible solutions that do not have to make use of introduced notions such as settled objectives. I will outline such a solution in Chapter 2: Section 2.3.

---

<sup>21</sup> Bratman adopts this the notion of “settled objectives” from McCann (1991: 26).

#### *4.3. Knowledge-backed: the problem of the lack of belief cases*

In many cases, it seems that I do a thing intentionally, but I do not believe that I am in fact going to do it. Call these *Lacking Belief* cases. This was touched on in the case of lucky action, but the problem runs deeper than just in such instances. Let us begin with those accounts such as those put forward by Velleman and Setiya, where intention is taken as reducing to a certain type of belief. An important criticism of this reductionist project, famously raised by Davidson (2001: 92), runs as follows: in order to show that an agent could have an intention to A, or intentionally A, without an intention-unique belief<sup>22</sup> that she would A, Davidson asks us to imagine a man (intentionally) writing his will with the intention of ensuring his family's well-being. However, due to his current situation of financial distress, he does not believe that he will actually do so. Making the same point, Davidson asks us to imagine a man making copies with carbon paper. The man has the intention to make ten copies, but does not believe that he will succeed in doing so. However, if he was to produce ten copies we would consider all of them to have been made intentionally. If the belief that you will A is not necessary for intentionally A'ing, then it seems the reductionist project is in some danger.

We have already encountered a typical cognitivist's response to cases such as these when we looked at Setiya's response to lucky action cases in Section 3.3. Indeed, Davidson's copy-maker is in many ways analogous to the case where Brandon the basketball player takes the shot without believing that he will succeed. To recall, on his initial account an action is intentional if the agent believes that she is doing it, or is more confident of this than she would otherwise be, or is doing this action by doing some other action that she does believe she is doing. When applied to the case of the copy-maker, Setiya's account would deliver the intuitively plausible verdict that the agent did intentionally make ten copies. However, despite providing the correct result in the case of the copy-maker, this approach provides implausible results when applied to cases where the chances of the agent succeeding in an action are *very* slight. This was illustrated by the case of Brandon, where his chances of success were a million-to-one.

---

<sup>22</sup> *Intention-unique belief* is a theoretical placeholder for whatever type of belief a given account of intention deems as characteristic of intentional action.

As we saw, Setiya hopes to avoid this possibility by moving away from talk of an agent's beliefs about what they are doing, to talk of an agent's knowledge-how. He then holds that a given action, A, should count as intentional when the agent responsible for it knows how to A, or does A by doing something else that she does know how to do. Setiya takes knowledge-how to be best understood in terms of the agent's dispositions. Without re-treading the same path discussed in the foregoing section, this move allows him to keep the intuitive answer to the copy-maker and will-writer cases, and provides a better – though still not wholly satisfactory – answer to the very low success chance cases (such as Brandon's). I take this to be the most compelling account of intentional action currently on the market, and it would be further improved by a better account of what it means to “know how to X” than the (admittedly provisional) dispositional story that Setiya provides. However, I do not pursue such an account here.

This criticism generalises beyond cognitivist accounts to those like Heuer's, as though intention, or intentional activity, is not reduced to belief in these accounts, it remains the case that the agent must have a self-referential belief that she is A'ing in virtue of which she controls her activity in order for the activity to qualify as intentional. Heuer recognises the worry revealed by *Lacking Belief* cases, and responds by sketching a parallel case to the carbon paper copier, one where the agent seeks to make only seven copies. In this case, she points out, if the agent was to make ten copies we would not think that the additional three were made unintentionally. Yet the only difference between the two cases lies in the respective agents' mental states – in their intentions and beliefs. Indeed, she claims that: “[i]n the original case he must believe that he is pressing so very hard on the page that it is at least possible for him to make ten copies; in the revised case, he must believe that he is pressing hard enough to make seven copies” (2014: 300).

However, this response is inadequate. There is a difference between believing that you are A'ing and believing that you are *possibly* A'ing. According to her account, in order for an action A to be intentional it must follow in virtue of a belief that the agent is A'ing. But in this case if the belief the agent has is that he is “possibly A'ing”, then we should describe the intentional action as “the agent is intentionally possibly making 10 copies.” This seems like a very strange thing to say. Rather we would be

inclined to say that the agent has the belief that he will *try* to A, and that he is intentionally *trying* to make 10 copies. However, this route is not open to Heuer given that she wishes to claim “kinship” (2014: 299) between her way of understanding intentional action and that of those who follow Anscombe in taking intentional action to involve a kind of non-inferential knowledge of what the agent is doing, she is committed to the idea that “[w]hen someone is acting intentionally, there must be something he is doing intentionally, *not merely trying to do*, in the belief that he is doing it [my emphasis]” (2014: 299). She could forsake this kinship, and I think she would be well-advised to do so, in which case she could endorse the solution of taking the relevant self-referential belief to be a belief to try. I will develop this insight in my own account of the role of belief in intentional action in Chapter 2: Section 1.3.

### **Concluding remarks**

In this chapter I have examined the features that we can expect to find in control accounts of intentional action: that they must reflect such action’s responsiveness to reasons as well as account for the epistemic conditions and requirements on intention and intentional action. Having provided this exposition, I then introduced the three broad approaches to understanding intentional action: reason-backed, intention-backed, and knowledge-backed accounts. With the descriptive component of the chapter complete, I then tested the three types of accounts with a number of cases of intentional actions at the fringes – cases where it is difficult to determine whether or not positive status is indeed justified. It was shown that none of the three were able to provide wholly plausible answers in all cases, though knowledge-backed accounts did fare best.

This done, I then presented targeted criticisms aimed at each approach in turn. For reason-backed accounts I argued that they struggle to provide a reasonable answer to cases of intentional akrasia without committing themselves to the view that it is impossible for an agent to act against the balance of all their reasons (be they epistemically transparent to the agent or not). Turning to intention-backed accounts, I showed how they struggle to explain the problem of mutually exclusive intentions without recourse to the introduction of theoretically undesirable notions such as

*settled objectives.* Finally, I discussed *Lacking Belief* cases, and some possible responses to them from knowledge-backed accounts. It was found that though conventional accounts of this sort struggle to accommodate them, there remains a strong case to be made for a knowledge-backed account that posits knowledge-how as the condition on intentional action. Though the knowledge-how account discussed, that of Setiya, still had a shortcoming in dealing with cases of actions with a very high likelihood of failure (such as lottery cases, or Brandon's case) I take it to be the most convincing account of intentional action considered here. It captures the most of the relevant phenomena, while still maintaining a strong theoretical consistency. For this reason, I take knowledge-how knowledge-backed accounts to be the most serious competitor to my own account.

In the next chapter I introduce and develop this account, which I believe can both make sense of the different fringes cases presented, avoid the various targeted criticisms applicable to the three approaches I have discussed, and prove more convincing than my chief competitor.

## CHAPTER 2: MY CONTROL ACCOUNT OF INTENTIONAL ACTIVITY AND INTENTIONAL OUTCOMES

### Introduction

Having examined some of the more prominent alternative accounts of intentional activity, as well as some of their shortcomings, in this section I lay out my own control account of intentional action – *System 2 Oversight* – which I take to evade these pitfalls. This account is a naturalised control account, where the control in question is understood as *oversight* by a *reasons-responsive cognitive mechanism*. Additionally, I contend that a necessary component of this form of control is the presence in an agent’s belief box of a *belief that she will try to X*, where X is the activity in question. I will also argue that this is primarily an account of *intentional activity*, rather than intentional action, and that there are good reasons for favouring this approach.

Of the three alternatives discussed in the previous section, my own account can be most easily understood as a reasons-backed account, though it has a few anomalous features that I take to give it an advantage not only over reason-backed accounts, but over all other accounts of intentional action. Though most of these features have extant histories, and so are not wholly new contributions, what I take to be noteworthy about the approach I sketch here is the way in which these features are deployed and brought together. Some of these features originate from other discourses in the philosophical tradition (such as *reasons-responsiveness*, which is drawn from the moral responsibility literature) or from the field of cognitive science (such as the *Dual Process Theory* model of human reasoning, which will play a significant role in my account). Others are rooted in the discussion surrounding intentional agency, but have not been used in the way I do here (most significantly the shift from talk of intentional action to *intentional activity*). Hopefully this originality in application can help move us toward a more complete account of intentional action – or rather, intentional activity.

It may be surprising that my own account would bear, on the surface at least, the most similarity to a reason-backed account. In recent years, it has been intention- and

knowledge-backed accounts that have tended to dominate the discussion. This is unsurprising, as not only do reason-backed accounts seem to fall short in regards to each of the varieties of fringe cases we have discussed – whereas the other accounts fare somewhat better – but they also open the door to issues involving the possibility of acting intentionally against one’s rational judgement. However, my account may be most similar to a reason-backed account, but it would be more accurately described as a *reason-responsiveness-backed* account, and most accurately as a *control-backed* account. What matters on my account is not the question of whether or not a given action was undertaken for a reason, but whether the agent had control over the action in question, i.e. whether a certain control condition is met, where this control condition does involve the *capacity* for guidance in the light of reasons. I will argue that making this move allows my account to avoid the problems of intending against one’s rational judgement.

Another facet of my approach that I take it to be a matter of significance is the central role played by the appropriate understanding of the actual *mechanism* that exerts intention-unique control. In order to identify this mechanism, I draw on insights from recent work in cognitive science, specifically those pertaining to Dual Process Theory. This focus on the actual mechanism gives us some insight into how the various features that clearly play such crucial roles in intentional action (such as reasons and beliefs), all come together. Recall IST:

**Intentional Status Transmission (IST):** An agent’s action or omission is intentional iff the agent possessed the right kind of control over it.

Where the “right kind of control” is replaced by intention-unique control. My account is an attempt to fill this placeholder by focussing on the *intention-unique control mechanism*.

I will explain my overall account in a series of steps: first I outline what I take to be the requirements that any account of the intention-unique control mechanism must meet to be convincing. This done, I then introduce and discuss Dual Process Theory (DPT), and based on insights drawn from here I put forward my own control condition on intentional action: *System 2 Oversight*. This condition will be incomplete

at this stage, and I will further refine it by defending the necessary role played by the belief to try in accounting for the epistemic component of intentional action. I then widen the scope of System 2 Oversight beyond solely intentional action, instead taking *intentional activity* to be the fundamental example of intentional behaviour. Building on this, I then discuss intentional effects and consequences, and how the intentional status of these can also be determined by understanding their relationship to System 2 Oversight.

Having thus presented my own account, I lastly demonstrate that an additional benefit that it brings is that it allows us to retain the unity of the three applications of the concept intention famously introduced by Anscombe (1963: 1): intentional action, intention-with-which, and intention-for-the-future. And that given my account of this unity, it is possible to provide resolutions to the problem of mutually exclusive intentions.

## 1. System 2 Oversight

### 1.1. Requirements of the intention-unique control mechanism

As has been identified in Chapter 1: Section 1.1., the intention-unique control mechanism must include the ability to guide activities in the light of reasons. I take this guidance in the light of reasons to be best understood as reasons-responsiveness, a notion first developed by Fischer and Ravizza (1998) in the context of their compatibilist account of moral responsibility. Fischer<sup>23</sup> presents two sorts of control, namely: (1) “guidance control” (2007: 56), and (2) “regulative control” (*Ibid.*: 57). Fischer illustrates the difference between these two sorts of control using the example of an agent driving a car and pulling off to the right in order to enter the parking bay of a coffee house. In this example, he states: “[h]ere you have a certain distinctive kind of control of the car’s movements – you have ‘guidance control’ of the car’s going to the right.” He continues to clarify that this “distinctive” sort of control can be differentiated from a case where, for example, the driver has an epileptic seizure, and collapses on the wheel forcing the car right. Fischer (2007: 78) describes this

---

<sup>23</sup> Though Ravizza’s contribution is invaluable to this discussion, as it has been Fischer who has carried these arguments forward, so I will at times be speaking of these as Fischer’s ideas.

guidance control as having “two chief elements: the actual-sequence mechanism that issues in action must be the ‘agent’s own,’ and it must be appropriately ‘reasons-responsive.’”

Regulative control, in comparison, is a quite different animal. This is the sort of control that follows from the agent having robust alternative possibilities to have acted other than how she actually acts. As Fischer (2007: 56-57) says:

Supposing that there are no ‘special’ factors at work – that is, no special psychological impairments, brain lesions, neurological disorders, causal determination, and so forth – and imagining (as above) that the car’s steering apparatus is not broken, you had it in your power (just prior to your actual decision to turn to the right) to continue going straight ahead, or to turn to the left, and so forth...you presumably (and apart from special assumptions) possessed freedom to choose and do otherwise: you had ‘regulative control’ over the car’s movements.

Setting aside regulative control, what does it entail for a mechanism to be reasons-responsive? In the simplest terms, it means that the mechanism is *receptive* enough to reasons to acknowledge their normative force, and *reactive* enough to these same reasons to adjust behaviour accordingly (Bratman, 2000: 454). Importantly, this does not mean that the agent must necessarily have the belief that a given consideration in favour of acting is a normative reason to act, but that the agent recognises the “call to action”, or the normative force of the consideration. This description still leaves open the task of specifying the scope of reasons to which a morally responsible agent’s actual-sequence mechanism must be responsive. As McKenna (2009) points out, if the mechanism is required to be too responsive to reasons then it would mean that an agent who acts immorally, and who knows that there are moral reasons not to, would not count as possessing guidance control – would not be considered to have acted freely. If the mechanism’s reasons-responsiveness is too weak, then it would allow “a person with only a very limited or insane pattern of sensitivity to reasons to count as satisfying the freedom condition.” In light of these problems, Fischer settles for moderate reasons-responsiveness. This kind of reasons-responsiveness requires that the mechanism must be receptive to a *significant* number of reasons (including moral

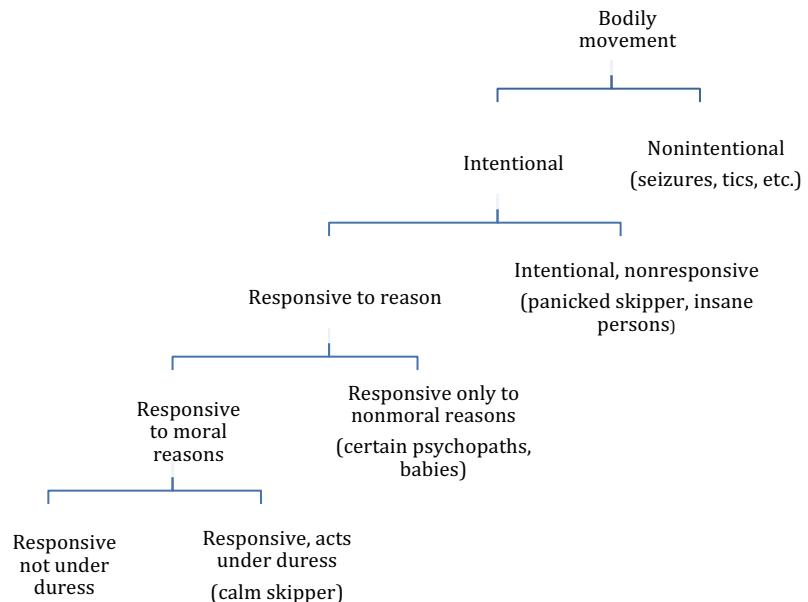
reasons) and must be reactive enough to adjust behaviour in light of, and in accordance with, *at least some* of these reasons (Bratman, 2000: 454). But given that guidance control cannot require access to alternative possibilities, how is this to be understood? Fischer's response is to employ counterfactuals of a certain sort. He argues that guidance control does not require that the actual-sequence mechanism had to be adjusted in light of the actual reasons present, only that the mechanism be reactive to reasons to the degree where in a nearby possible world some (relevant) reason – though not the reason that was actually ignored – could have resulted in the behaviour being different from what actually happened.

I find this best expressed through an example, and so I will paraphrase one provided by McKenna (2009): imagine that Matilda is dancing at a party. Matilda enjoys dancing very much, and it would take a considerably powerful reason for her to adjust her behaviour. This said, she is still moderately reasons-responsive as she would stop dancing if anybody offered her ten-thousand dollars or more, or if she was informed that if she stopped dancing twenty people's lives would be saved. Now imagine that Matilda is brought news that her mother has suffered a serious accident and needs her help. Matilda is receptive to this being a possible moral reason for her to stop dancing, so she appreciates its normative force. However, imagine that she does not stop dancing, despite recognising by her own notion of decent conduct that she should stop. We can imagine her saying, "I really should stop dancing to help my mom, but I'm having too much fun!" Clearly then, she is unreactive to this relevant moral reason to do otherwise. However, according to Fischer's argument, this does not absolve her of moral responsibility provided that this case of insensitivity to reasons is "situated" within a set of cases that demonstrate "a rich sensitivity to some rational and stable range of reasons." That is, as long as Matilda's mechanism is sufficiently reasons-responsive such that there are sufficient near possible worlds where she would adjust her behaviour in the light of relevant reasons to do so, *even though her mom needing her help will never be one of these*.<sup>24</sup>

---

<sup>24</sup>As McKenna (2009) notes, there is a certain "irony" in this final point. The very moral reason for which we seemingly hold Matilda morally responsible for not being reactive to is the very reason that she could never have been reactive to. McKenna notes, however, that it is not obvious whether this irony forms grounds on which to criticise Fischer's position. I would suggest that it does not.

Fischer's own use of reason-responsiveness is embedded in moral concerns – unsurprisingly given the context for which he developed it. This means that his version of reason-responsiveness involves a number of considerations that serve to capture features of moral responsibility practices. This is most significant when thinking about the proper scope of receptivity and reactivity. There are, after all, agents capable of acting intentionally, yet are not legitimate targets for moral responsibility, such as young children, those suffering from certain pathological conditions, and perhaps even animals. Fischer excludes these cases by proposing the need for an ownership condition,<sup>25</sup> which excludes many such cases, as well as arguing that these might represent instances of *non-reason-responsive intentional action*. Into this category he places examples such as certain kinds of insanity (or pathological conditions) as well as the case of a “sea captain who panics in a storm and is impelled to jettison his cargo by an irresistible fear” (1998: 82). This is a category of actions that my account cannot accommodate, and so something must be said about it. Fischer and Ravizza (*ibid.*: 83) provide the following diagram to illustrate the breakdown of the relationship between reason-responsiveness and intentional status as they see it:




---

<sup>25</sup> I will be discussing Fischer's ownership condition in Chapter 3, where I discuss his account in a moral responsibility context. It is sufficient here to note its existence.

On this classification, the problematic category is “Intentional, nonresponsive”. My first response to the possibility of this category is to reject the claim that the agent in the examples provided – that of the panicked skipper and the insane – are in fact wholly unresponsive to reasons, and furthermore if they are, then their actions would not be intentional. In the case of the panicked skipper, it is difficult to imagine that her fear was such that she was thoroughly unresponsive to reasons, though this responsiveness may be reduced (to varying levels of severity). However, if we accept the stipulation that the skipper’s fear is *truly irresistible*, such that her reason-responsive mechanism exerted (and was capable of exerting) no guidance over the action, then this does not seem to be an intentional action. The skipper could justifiably claim, “I didn’t throw the cargo over intentionally! I was overcome by fear and had lost control of myself!” She would be no different to the person emerging from a truly blind rage for the first time and claims of the harms done, “I didn’t do these things intentionally! I was not in control of myself!” Crucial to note, this does not necessarily absolve either the skipper or the person emerging from the rage from moral responsibility (as will be discussed in Chapter 3: Section 4.3.). I think that most (if not all) resistance to recognising the justifiability of the skipper making such a claim results from the implausibility that such a complex action as throwing the cargo overboard could be irresistibly impelled by fear, without any responsiveness to reasons. In the case of the insane, I take the same reasoning to apply. In most cases, such individuals are likely to retain some (though reduced) capacity for responsiveness to reasons, though in those cases where such capacity is in fact absent, they can justifiably deny the intentionality of their behaviour, as positive intentional status is indeed absent.

Where I am in agreement with Fischer, however, is that intention-unique control is not concerned uniquely, or indeed necessarily, with *moral reasons*. It is possible to imagine an *idealised* psychopath, an agent who is perfectly reasons-responsive to all egoistic and prudential reasons, but is pathologically – and entirely – incapable of responding to moral reasons qua moral reasons (they could, of course, learn to theoretically identify those considerations that others take to be moral reasons, but what they lack is the ability to recognise the special normative force of such reasons). Such an agent would most certainly be capable of acting intentionally, but would not meet Fischer’s criteria for moral responsibility as they are not “moral agents” (1998:

82). The final important difference is that intention-unique control does not require that there *be* a reason in the vicinity to which the agent could have responded by guiding her activity, but rather that the agent must have had the *capacity* to respond *if* such a reason were present. This is a kind of *oversight*, as when a driving instructor has oversight over his learner in a situation where he rides along with said learner and has the capacity to intervene if there is cause for it.

This talk of capacity is most easily understood counterfactually, though not in quite the same way as Fischer envisages. To say that the agent had the capacity for oversight is to say that in some close possible worlds the agent, through the actual-sequence mechanism, could have responded to reasons (moral or otherwise) pertaining to her activity and guided them accordingly. This responsiveness does not need to be consciously exercised. An agent could carry out an action and, in the moment, not be consciously aware of doing so, and yet the mechanism resulting in the action could be suitably reason-responsive. If I am driving down the road to the store and come to an intersection in the road, it is possible that I am so accustomed to the drive that I perform the turn toward the store without, in the moment, being consciously aware of doing so – my mind might be entertained in some speculation about the nature of control, and so distracted from my activities. What matters in determining whether the turning is under intention-unique control, and thus intentional, is whether or not, in that moment, I was reason-responsive such that in close possible worlds the mechanism that resulted in the action could have been adjusted in the light of salient reasons.

Returning to talk of the scope of responsiveness, some things are readily apparent: the degree of sensitivity to reasons – in terms of both receptivity and reactivity – necessary for positive intentional status is far lower than it is for moral responsibility. First off, and as already mentioned, no responsiveness to moral reasons is required for positive intentional status. And where non-moral reasons are concerned, we often think of agent's as acting intentionally who have very limited responsiveness to normative reasons to act – such as very young children – whereas we usually exempt them from moral responsibility. Consider also that if there was an agent who was sensitive to only one reason, then if that agent were to act upon that reason, or act in a context where that reason was salient, it would make sense (presuming certain further

conditions, such as epistemic ones, are met) that this action would be described as intentional, whereas it seems likely that such an agent would not be responsive enough to count as open to moral responsibility. However, it is also important to note that if this same agent performed some other behaviour where this is not the case, then the action would not count as intentional. This leads me to say that intention-unique control over a given action requires that the actual-sequence mechanism must be *minimally* receptive and reactive. By *minimally* here is meant that the mechanism must either have responded to a reason, or would be responsive to a reason in a minimally close possible world if there were such reasons present.

Importantly, this does not mean that an action can only be intentional if there is normative reason in the vicinity. It must just be such that *if* there has been such reasons the agent had the capacity to have responded. Also, worth noting is that an agent can be very reason-receptive but if she is not also somewhat reason-reactive then attributing positive intentional status would be incorrect – an agent like Bill the drugged vacationer from Chapter 1: Section 1.1. would be an example of this.

It can be asked, however, if this doesn't throw the net of intentional action too widely. To see this, let us again consider action slips. Action slips, to remind, occur when an agent falls into an action or chain of actions (usually habitual ones) without realising they are doing so, and often in opposition to her explicit intentions, and who respond when made aware of this by stating things like, “I didn’t know I was doing that!” or “I hadn’t realised what I was doing!” These cases seem to me to be ones where it would be implausible to conclude that the agent is acting intentionally, yet my account on the face of it would count them as such, since there is likely to be at least some reasons which, had they been in the vicinity, would have allowed the agent to adjust, i.e. stop, the behaviour. We can imagine the man who went to get his shopping bags from the shed but fell into the action slip of putting on his gardening gloves (as is his habit). Presumably, had his gloves been covered in spiders he would have responded and stopped his behaviour. So, given the description of reason-responsiveness I have given above, if he had completed his action slip then we could have to conclude that it was intentional.

To avoid this outcome, we must recall the epistemic condition on intentional action. The agent must have some kind of epistemic relationship to the action in order for it to count as intentional. Let us say that this state should be some belief, and call it an *intention-unique belief* for now. I will discuss precisely how this intention-unique belief should be fleshed out in section 1.3., but for now the question to ask is what role such a belief should play in the actual-sequence mechanism. It seems most plausible that the role that the belief plays, and the reason that it is necessary for intention-unique control, is that only if the actual-sequence mechanism responds to reasons in the light of this belief can the action in question really have been under reason-responsive control. This belief does not need to be conscious to play this role, it need only be in the agent's belief box, there to be accessed in the appropriate situations – as is the case with the vast majority of our beliefs. Usually the best way to identify if such a belief was present is if the agent can reflexively access it. In action slip cases, this is precisely what is missing. The agent has no accessible belief about doing the action in question, hence their surprised reactions when being made aware of them. The result of this is that the knowledge-backed proponents are correct on at least one count: an action can only be intentional if the agent had some appropriate belief about undertaking it. I will argue in a latter section that this belief is best understood as a *belief to try*.

It could be asked though, why does an actual-sequence mechanism with these features impart positive intentional status to an action? The answer to this is that the outcomes of such a mechanism reflect the agent's rational self. This is well expressed by Velleman (2014) when he states:

If a person's constitution includes a causal mechanism that has the function of basing his behavior on reasons, then that mechanism is, functionally speaking, the locus of his agency, and its control over his behavior amounts to his self-control, or autonomy.

This is not to say that all there is to the self is its rational faculties, there are doubtless over important aspects to the self, such as desires and beliefs. However, when it comes to agency, we identify – and are identified – uniquely with those actions that we undertake for normative reasons. And this we can only do if we possess some

appropriate mechanism, the same way that I can only see if I possess the appropriate perceptual mechanism.

This gives us our first two requirements that an intention-unique control mechanism must meet: (I) the results of such a mechanism must be at least minimally reason-responsive, and (II) that the mechanism must account for the role of intention-unique belief. There are a further two requirements that such a mechanism must meet, namely: (III) it must allow the possibility that an agent act – and act intentionally – against the balance of reasons, and (IV) it must be able to result in both intentions and intentional actions, and not necessarily together. An eagle-eyed reader would have noticed that (III) and (IV) are necessary in order to avoid two of the pitfalls that plagues the accounts of intentional action discussed in the previous chapter. (III) Serves to ward off the criticism directed at Davidson's account that, by requiring that all intentional actions are caused and rationalised by all-things-considered judgements, it was impossible for an agent to act against the balance of reasons (both those that are epistemically transparent and epistemically opaque to the agent). This is a very implausible way to think about human action, it certainly seems that we can act against the balance of reasons, and that when we do so this it is not because there were some opaque reasons at work. The purpose of (IV) is to avoid commitment to the idea that every intentional action must have some related intention (or every intention needing an action to eventuate), and to partly avoid the danger of holding mutually exclusive intentions – though even if (IV) were to be met, this would not rule out the possibility of mutually exclusive intentions. It is however a necessary piece of the solution.

Mutually exclusive intentions would run the risk of being very frequent were it the case that every intentional action A had to be backed by an intention to A. To recall Bratman's gamer example, we would be compelled to conclude that the gamer had the intention to strike T1 and the intention to strike T2, and so be in a contradiction. However, as we saw in the discussion in Chapter 1: Section 4.2., there may still be cases where an agent does hold an intention to A and an intention to B, and where A and B are mutually exclusive. To resolve the danger permanently, it is necessary to slightly rethink the rational requirements characteristic of intentions. I will discuss

these requirements more in the next section, however my final resolution of the problem of mutually exclusive intentions is presented in Section 2.3.

This leaves us with a very fundamental question to ask: what is this mechanism? Fischer (2007: 78-79) purposefully avoids discussing the details of what this mechanism could be, keeping his own description vague: “a process or a ‘way things go’ along the path that leads to the behaviour in question,” and even explicitly rejects the identification of the mechanism with practical reason (Fischer and Ravizza, 1998: 86). Velleman (2014), though not strictly speaking of intentional action but rather what he calls *autonomous action*, provides a more determinate answer and does identify it with practical reason. Without wishing to engage in a protracted debate concerning the nature of practical reasoning, my view here is that the intention-unique control mechanism must be open to oversight by practical reasoning, even if, in a given instance of the mechanism’s operation, no such reasoning took place. But I wish to be more specific about the identity of the actual mechanism at work. The hope in doing so is that by properly elucidating the properties of this mechanism that plays this role in our lives, it may be possible to glean insights into other facets of intentional agency. I turn to this attempt in the succeeding section.

### *1.2. Dual Process Theory and System 2 Oversight*

I take it to be a theoretic virtue for any account of intentional activity to be compatible and consistent with our current best theories in the natural and cognitive sciences, i.e. if it were naturalised. An account of intentional activity that fails to do this finds itself in the unenviable position of having to either provide an error theory for those findings in the sciences with which it is inconsistent, or to “damn the data.” Though the former is certainly the preferable option between the two, barring there being some substantial theoretic virtue that would be lost in naturalization, neither should be preferred over the provision of a naturalised account.<sup>26</sup> I take the control account I am

---

<sup>26</sup> It can justifiably be asked: “what do I mean by naturalised?” A helpful answer to this question can be found in Roughley (2004). In this piece, he lays out several different uses of the term “naturalist” (I assume that “naturalist” account is synonymous with “naturalised” account, for the purposes of this paper) in philosophy, labelling them N1 – N3 (2004: 51-52):

(N1) A philosophical approach is naturalist iff its procedures are consistent with the assumption that its subject matter has come into being as a result of evolutionary processes.

developing in this chapter to be such a naturalised account as it builds directly on insights drawn from some of our leading theories in the field of cognitive science.

Whereas distinguishing intentional activity from the rest of human behaviour has been a central preoccupation in the philosophical discourse surrounding action, in the cognitive sciences the focus has tended to be more on distinguishing automatic activities from consciously controlled activities (Norman and Shallice, 1986; Kahneman, 2011; Moors and De Houwer, 2007). As was mentioned in Chapter 1: Section 3.2., automatic activities are distinguished from mere events (such as slipping on the floor, or being carried away by the wind) by the fact that they still fall under some kind of control, though that control is clearly different to conscious or online control, and presumably different (at least in some cases) from intention-unique control, given that such automatic actions are often characterized as unconscious and unintentional. This latter fact means that these theories have the obvious potential to have bearing on the discussion surrounding intentional activity.

A fairly recent theory of human cognition, currently influential in cognitive science, and the one I will be primarily leaning on, is Dual Process Theory (DPT). The basic suggestion of DPT is that human cognition can be roughly divided into two sets of cognitive systems – System 1 and System 2 – where the operation of System 1 is thought to be typically fast, automatic, and energy inexpensive, and System 2 is typically slow, consciously controlled, energy intensive, and has the capacity to be responsive to the largest set of reasons in a given context (Evans and Frankish, 2009; Kahneman, 2011). Rather than a single unitary system, System 1 is usually understood as being a classification that covers a large number of individual processes that have the characteristics attributed above, and these processes are thought to be massively parallel, which facilitates the typical speed and low energy costs associated

---

(N2) A philosophical approach is naturalist iff its procedures are compatible with those of the natural sciences.

(N3) A philosophical approach is naturalist iff its aims and procedures are aims and procedures of the natural sciences.

He further points out that a given philosophical approach might well count as naturalised under some of these uses, but not under others. I will be accepting these as formulations of the different conditions an account may meet to be deemed naturalist. My own account I take to meet N1 and N2, and so I hold that it can be justifiably labelled a naturalised account.

with them. At least System 2, which is taken to be unitary, is taken to be reasons-responsive to normative reasons, and there is ongoing debate about whether or not System 1 processes responds to normative reasons in a merely associative and heuristic manner, or whether such processes might involve more goal-directed reasoning in their operations (for the former view see Kahneman, 2011, for the latter see Carruthers, 2014). What is important is that this sensitivity in System 1 is never taken to be a sensitivity to reasons qua normative reasons. Evans and Over, for example, are on one end of the spectrum in suggesting that System 1 processes can indeed be rational, but they differentiate between two different kinds of rationality, namely (1996: 8):

Rationality<sub>1</sub>: Thinking, speaking, reasoning, making a decision, or acting in a way that is generally reliable and efficient for achieving one's goals.

Rationality<sub>2</sub>: Thinking, speaking, reasoning, making a decision, or acting when one has a reason for what one does sanctioned by a normative theory.

System 1 processes, Evans and Over contend, can meet the demands of Rationality<sub>1</sub>, but not those of Rationality<sub>2</sub>. Carruthers (2014: 9) describes this distinction as one between “ecological” and “normative” rationality, with System 2 (or what Carruthers calls *Reflective Reasoning*)<sup>27</sup> being uniquely capable of the latter. He attributes this feature to the fact that the operations of reflective reasoning are “action-based” (*ibid.*: 4), meaning that it is open to guidance by things like verbal instruction and normative beliefs.<sup>28</sup> By action-based is not meant that such reasoning necessarily culminates in

---

<sup>27</sup> Though Carruthers (2014) sees his dichotomy between intuitive reasoning and reflective reasoning as a replacement for talk of System 1 and System 2, since he takes the opposing list of features typically associated with each system to fail at capturing a real distinction. However, despite this, it is equally easy to interpret what he is doing as simply providing a different account of the two systems. And since, as far as I can see, nothing in my argument hangs on this, I will be treating reflective reasoning and intuitive reasoning as Carruthers’s versions of System 2 and System 1 respectively.

<sup>28</sup> Carruthers also takes the action-based nature of reflective reasoning to mean that it must be open to intentional control. This is plausible, after all we do sometimes intentionally think or reason through a problem. However, unlike Carruthers I do not think that we should say in these cases that the reflective reasoning process is intentionally controlled, but rather that any case of thinking or reasoning that is undertaken by such a process is, by definition, intentional. The reason for this difference is that I think Carruthers associates intentional control with full-on conscious control, which we have already established is too strict a control condition on intentional action. Rather, the control condition involves oversight by a mechanism that is reflective of the agent’s reason-responsiveness to normative reasons, and such a mechanism is always in oversight when we reason using a reflective reasoning process (or System 2), as it is just such a mechanism. Put differently: my thoughts and judgements that result from

overt action, though in at least some instances this will be the case. But on many occasions the operation of reflective reasoning will instead yield (individually or in combinations) mental rehearsals of action, judgements, beliefs, and/or intentions. However, since my interest is on the question of moral responsibility's relationship to intentional activity, I will be focussing on those operations of the systems that yield intentions and actions.

The fast/slow and energy inexpensive/energy expensive distinctions are simple enough to grasp, but it is worth taking a moment to consider the final distinction: automatic/consciously controlled, which has the most relevance for my arguments regarding intentional activity. There are various different definitions of automaticity when applied to cognitive processes, some requiring that these processes not be goal-directed (Carruthers, 2014: 13), others that it be generally inaccessible (Sun, Lane, and Matthews, 2009: 256), and yet others claim that a process is autonomous when it can “control behaviour directly without need for any kind of controlled attention” (Evans, 2009: 42).<sup>29</sup> For my purposes I will follow the lead of Norman and Shallice (1986: 1-2) in understanding an automatic process as a process that can operate, perhaps instigating activity, without the awareness of its performance. It is important to stress that what the agent is necessarily unaware of here is the cognitive process that results in a given activity, not necessarily the resultant activity itself. Whether or not the agent is aware of the actual activity is a further question (and an important one, as it will turn out, for determining intentional status). By contrast, a process that is consciously controlled is taken to be influenced by verbal instruction and normative beliefs, “globally broadcast” (Carruthers, 2014: 7-9), and linked to an agent’s sense of agency. It is also assumed that such a process is in principle available to the widest set of rational tools that the agent can bring to bear. It is worth noting that the activities that result from System 2 processes can necessarily be assumed to be under agential control, as such processes and the activities they result in are both open to guidance from beliefs, as well as modifiable in response to normative reasons.

---

my reflective reasoning is always the result of a reason-responsive mechanism with oversight over these results.

<sup>29</sup> To get a sense for the complexity surrounding the concept of automaticity see Moors and De Houwer (2007) for an excellent, and comprehensive, overview.

In the most usual version of the DPT model, the *default-interventionist* version, most of an agent's activities in their daily life are directed by System 1 processes, with System 2 only intervening or being called upon when the System 1 processes breakdown or new input – in the form of changes in the context that yield new considerations to act that may count as reasons for the agent in question – is received. When this occurs, System 2 can intervene to adjust the agent's behaviour. An easy example is that of a man, Tom, making tea: we stipulate that his behaviour is being directed by System 1 processes, as he is actively thinking about something else. As part of the routine of making tea, Tom reaches up to a shelf where the sugar is usually kept. However, as it happens, Tom's wife used the last of the sugar for her own tea, and had not replaced it, leaving Tom's hand to close around empty air. When this occurs, Tom's System 2 is called in to intervene and determine a solution. So, when the context changed, the System 1 process(es) at work, possessing only a limited capacity to respond, failed to provide the appropriate outcome, and System 2 – which had oversight over the events occurring – activated in response. Crucially, it does not require a failure of a System 1 process to achieve the expected result in order for System 2 to intervene, sometimes this occurs when new considerations enter the picture. This time Tom is driving to the store to pick up some more sugar. He is angrily thinking about his wife's negligence in not replacing it herself, and his driving is being directed by System 1 processes, this being a long, straight, and boring stretch of road. Then, suddenly, a deer appears on the road. Suddenly Tom's awareness snaps to the road and the deer and he slams the breaks and pulls the steering wheel to the side. The explanation for this from DPT is that System 2 intervened when a new reason for acting was recognised in the vicinity.

There are two features of System 2 that need more expansive explanation: that of *global broadcasting* and of the relationship to the *sense of agency*. Global broadcasting theory is a cognitive architecture intended to explain the coincidence of the broadcasting of outputs of the ventral-temporal visual system with the conscious experience. The gist is that this global broadcasting makes the content of the broadcast available to a vast array (some, like Carruthers, 2014, argue the full suite) of unconscious processes. I take these unconscious processes to be System 1 processes. There are then two possible stories for System 2's relationship to global broadcasting: the first is that there is some particular cognitive process, called System

2, that once engaged has the ability to “call on” the array of System 1 process, and that global broadcasting is part of this ability. The second story argues that there is no particular process called System 2 to be identified, rather what we take to be the workings of System 2 are in fact the repeated cycles of the operation of otherwise unconscious and unreflective forms of cognition, and that these cycles require global broadcasting. I take my view to be compatible with either of these interpretations, though I do favour the latter, as it does seem more evolutionarily likely that such a system of broadcasted repetitions of existing processes would develop, rather than an entirely new one. What is important for the purposes of my account though, is that the operation of System 2, however understood, brings its content to consciousness.

Turning the sense of agency. This refers to an agent’s subjective experience of controlling her actions and through them events in the external world. As described by Haggard and Tsarkis (2009: 242):

As we perform actions in our daily lives, we have a coherent experience of a seemingly simple fluent flow from our thoughts, to our body movements, to the effects produced in the world. I want to have something to eat, I go to the kitchen, I eat a piece of bread. We have a single experience of agency—of control over these events—because our cognitive representations of the successive stages of sensorimotor control are tightly linked together.

Importantly, this has to do with the *subjective experience* of agency, not agency itself. It is possible that an agent might experience agency over some outcome that was in fact outside of her control, and this has been borne out in a number of studies.<sup>30</sup> The reason this is of interest to my account, is that the linkage mechanism that is most commonly assigned the role of connecting up the successive stages of sensorimotor control – a process called *intentional binding* – is *action prediction*. Agents experience an action as theirs when their prediction of the outcome matches closely enough with the actual outcome brought about (including temporal factors, so proximity in time is relevant). This action prediction must be conscious in order for intentional binding to take place. This is the connection with System 2: these action predictions are results of System 2 – indeed on one influential understanding of

---

<sup>30</sup> For examples see Farrer et al., 2003; Metcalfe and Greene, 2007; Sato and Yasuda, 2005.

System 2, all that this system's operation brings about is *conscious action rehearsal* that then can result in the implementation of the action by various System 1 processes (Carruthers, 2014), which maps on well to the role ascribed to action prediction in the sense of agency literature.

Does this mean that every action over which the agent has an experience of agency is intentional? Certainly not. As the afore-mentioned studies have shown, our experience of agency can come apart from the actual presence of agency. This occurs because the action in question is not in fact amenable to adjustment in the light of reasons, i.e. is not an action that is under System 2 Oversight, but followed from some action that was, and which was predicted to have the given effect or consequence. Despite this fact, I take the relationship between the operation of System 2 and the sense of agency to provide circumstantial evidence for the idea that our intentional agency is exercised through the operation of System 2, since intentional binding is *usually* an accurate indicator for the presence of agency, and that the times that it fails can be adequately explained from within the Dual Process Theory account, since there is nothing in the DPT account that requires that an agent always has correct beliefs regarding the consequences of her intentional actions.

Given this overview, I now wish to propose a control account of intentional action that is informed by – and is compatible with – insights into human activity drawn from Dual Process Theory. I will be relying primarily on the default-interventionist interpretation of DPT, because it remains the most common interpretation, though it has spawned a number of variants (Stanovich, 2009; Evans, 2009; Carruthers, 2014). Under my account, the type of control that unifies all cases of intentional action is *System 2 Oversight (S2O)*, which at this stage can be provisionally understood as follows:

**System 2 Oversight:** A given *action*<sup>31</sup> X, undertaken at time t, is under S2O, and so intentional, iff:

- (1) At time t the agent X'd

---

<sup>31</sup> It is important to stress that this condition as provided here accounts only for intentional actions, and not yet for intentional omissions, effects, or consequences. I turn to broadening the account in the succeeding sections of this chapter.

- (2) (a) X is under the guidance of the operation of System 2 or (b) is under the guidance of System 1 and System 2 had oversight
- (3) The agent's System 2 operation in this case is at least minimally responsive
- (4) The agent had an intention-unique belief pertaining to X

Condition (1) being fairly self-explanatory, I will pass it over and consider each of the succeeding conditions in turn. (2) captures the origination of the action from the agent's "locus of agency", as Velleman might call it. To meet (2) an action can result from either a System 1 process, or have followed from the operation of System 2 (or from a System 2 process, if this is your preferred description), or it can be partially the result of one and partially the result of the other. This kind of mixed action can easily be found if we turn back to think about automatic action, particularly over-learned ones. When I perform a tennis shot, my action is a result of the operation of System 2 in so far as I consciously reasoned about the placement of the shot, and how this fits into my overall strategy for this point. But at the same time, I am also intentionally doing an entire array of things that are not being run through my System 2, but by System 1 processes – my footwork, my grip movement, etc. It is not enough, however, to be the result of the workings of the two systems. What is needed is that the given action was under *guidance*, meaning that the agent would have been able to adjust the given action in the light of the reasons that she is responsive to.

Having subconditions (a) and (b) ensures that this account captures the feature of intentional action that has been stressed by most thinkers since Anscombe, that intentional action must be action that can be open to demands for explanation in terms of the agent's responsiveness to normative reasons. This will be unproblematically the case for actions that meet (a). However, as we saw, this does not mean that what is required is that there always be a normative reason that the agent must be acting on, but rather than the agent have the capacity to adjust her behaviour in the light of reasons. Therefore, what is required for positive intentional status is not that an action result from the operation of System 2, but rather that System 2 must have *oversight*, as understood in the default-interventionist picture of DPT. Ergo an action can be intentional if it meets (b), rather than (a).

Condition (3) is necessary since the scope of responsiveness can be varied. The reasons for choosing minimal reason-responsiveness was discussed in Section 1.1.

Finally, (4) is there to account for the epistemic condition on intentional action. It is not enough that an action, A, be under System 2 oversight, it is also necessary that the agent have at least the belief that they are trying to A. Since the content of System 2's operations are globally broadcast, it follows that when such operations result in action, the agent will have the necessary belief. On the other hand, this will not always be the case for actions that are under oversight. Again, consider something like an action slip, which is characterised as a case where the agent does not have a belief as to what he's doing until it's brought to his attention, and even then, he has no belief to recall, but rather learns about his behaviour observationally. This said, there is not a requirement that the belief must be conscious at the time of acting – indeed, such a requirement would render large numbers of clearly intentional, System 1-backed, actions unintentional. I explain and defend my use of a belief to try in Section 1.5.

So how does this account fare in meeting the four requirements outlined in the previous section? The first requirement is met at face value, since System 2 oversight is the cognitive mechanism that enables us to be reason-responsive. Requirement (II), that the intention-unique control mechanism must account for the role of intention-unique belief, is met by the addition of condition (4). Requirement (III) allows for cases where agents act intentionally against the balance of reasons. This is possible since oversight does not mean that System 2 will *always* intervene in System 1 operations, even if there is a normative reason to do so. It seems likely that many System 1 processes are sensitive to an agent's desires and attitudes toward some end, even ends that run counter to what the agent takes herself to have normative reason to pursue, and that actions that result from this might be under System 2 Oversight, but in any particular instance it is possible that System 2 does not intervene. So, for example, if I judge that smoking the next cigarette is not the right course, but my desires push me to an intention to smoke the next cigarette regardless, this smoking would still be intentional provided that the agent had the capacity to respond to reasons not to do so. Capacity, to recall, being understood here in a Fischer-style counterfactual manner. What is important to note though, is that if the agent did not

have the capacity to adjust the behaviour in the light of reasons, had there been salient ones in the vicinity, then the action would not be intentional. In this version of the case, the agent's behaviour is too compulsive to count as intentional, and is more like the case of the kleptomaniac, or of an addict who is incapable of resisting his addiction – such cases may not only be unintentional, but they also seem to be involuntary. There is a similarity here to Heuer's point from Chapter 1: Section 1.3., that what matters in determining whether or not an action counts as intentional is not that the action be done for a reason, but that the action be under control in such a way that it could be stopped, modified, or continued, in the light of salient reasons. Even when an action (or judgement) is the result of System 2 – not merely under oversight – it is still possible that this action (or judgement) could run against the balance of reasons. Though System 2 brings to bear the largest rational toolkit and is responsive to normative beliefs about reasoning, nothing in the DPT account requires that it always yield outcomes in line with the balance of reasons. Put another way, judgements resulting from System 2 need not be thought of as all-things-considered judgements in the Davidsonian sense, and so actions following from such judgements can be akratic.

Finally, requirement (IV), that the mechanism must be able to result in both intentions and intentional actions and not necessarily together. To recall, this requirement is necessary to avoid the dangers of mutually exclusive intentions. Fortunately, though it does not resolve the problem entirely, System 2 Oversight can easily explain the possibility of intentional actions without intentions, since either can be the result of a System 1 process or the operation of System 2, though, being intentional, both must be under System 2 Oversight. This is sufficient to account for cases of intentional action with no intention, like the automatic actions discussed in Chapter 1: Section 3.1. However, it is not yet enough to evade the problem of mutually exclusive intentions entirely.

I take this account to be unlike reason-backed, intention-backed, and knowledge-backed accounts, in that the presence of *control* itself is taken to be the characteristic feature of intentional action. What is more, this control is exerted through a mechanism. That the operation of this control mechanism might involve a role for reason, intentions, and beliefs, does not alter the direction of priority. In the

discussion of the fringe cases in Chapter 1, it was shown that there are cases of intentional action where there appears to be no reason or intention involved, and with all three of these features it is possible that an agent holds them but does not in fact act intentionally – they are not *sufficient* for positive intentional status. In contrast, System 2 Oversight is both necessary and sufficient.

This then gives us what I take to be the backbone of a compelling formulation of intention-unique control for intentional action: System 2 Oversight. However, this is not yet a complete picture, as it has not yet fully defended my version of the epistemic condition on intentional action, and not wholly resolved the problem of mutually exclusive intentions. It is also not yet a full account of intentional *activity*, as it excludes omissions. Nor does it sufficiently capture the intentional effects and consequences of such activity. To see this easily, consider the case of Aaron the assassin. We do not want to say that Aaron intentionally killed his target, though he had the intention to, since he killed him by shooting him in the head, whereas Aaron believed that the only way to kill his target was by hitting him in the heart and missed. On the account I have laid out, Aaron’s killing of the target is clearly intentional: the action is the result of a proper cognitive mechanism, System 2 had oversight, System 2 was minimally reason-responsive. My response to this is that the control condition for intentional actions and omissions is best understood as a single condition on intentional activity, whereas intentional effects, and intentional consequences have a differently formulated control condition – involving a different intention-unique belief, though the same notion of control is central to it. I will be tackling these remaining elements in the following order: I first unpack my approach to intention-unique belief in Section 1.3. Secondly, I expand my account to include all intentional activity in Section 1.4., and intentional effects and consequences in Section 1.5. Finally, I provide my full solution to the problem of mutually exclusive intentions in Section 2.3.

### *1.3. The belief to try*

In this section I argue for what Yair Levy (2017: 1) might call “Weak Intention Cognitivism”: the position that any intentional action requires a related belief. To clarify my views, I do not endorse the strong and controversial position that practical

reasoning generally, or the rational requirements on intention particularly, are reducible to the demands of theoretical reasoning. Nor do I claim that intention itself is best understood as a belief. What I will defend is the far more modest claim that whenever there is an instance of intentional action, A, the agent necessarily has an intention-unique belief, and this belief is best understood as a belief that she or he will try to A.

As we saw in Chapter 1, knowledge-backed accounts are perhaps the strongest rivals to the account that I am developing here. However, one of the prominent criticisms of such accounts has been that agents sometimes A intentionally without believing that they will A. It was also suggested in that discussion that perhaps what is required is not a belief that the agent will A, but a belief that the agent will *try* to A. However, this move faces stiff objections. We have already been introduced to Davidson's will writer and copy-maker counterexamples. I argue that in both these cases the agent in question can be said to have a belief that he or she will try to A, where A is that action in question (I will be calling this belief, a *belief to try*). The will writer believes that he is trying to secure his family's future, the copy-maker believes that he is trying to make ten copies. However, more recently Holton (2009: 22), conceding that the belief to try approach can resolve the Davidsonian counterexamples, puts forward another case, the Forgetful Cyclist, which he argues cannot be resolved using a belief to try. He presents the case as follows

You have some library books that are badly overdue; in fact so badly overdue that your borrowing privileges are about to be suspended (a major inconvenience) if you do not return or renew them by the end of the day. Since you have finished with them, the best thing would be to drop them off at the library on your way home; but that is after the departmental seminar, and you know that, once you get on your bicycle with your head full of ideas from the discussion, you are all too likely to cycle straight home. Alternatively, you could renew them online; but that would require your library password, which is scribbled on a piece of paper lying somewhere on your desk at home. If you renewed them online you would not need to take the books home with you, but you need to take your laptop, which you would otherwise leave at work. In the end you head for the seminar with your bag weighed down by both the library

books *and* your laptop, moved by the thought that you will avoid suspension one way or another.

His contention is that this case is resilient against intention-unique beliefs as beliefs to try explanations. He takes it as obvious that if the agent described was to fail in returning the books because she forgot about doing so post the departmental seminar, she would not have “even tried to take them back”. The idea seems to be that because the reason that the agent failed to fulfil her intention was due to forgetfulness, this means that the agent did not even try. Interestingly, Holton implies that had the agent doubted she would succeed in returning the books because the library was, for example, “under siege,” then this would somehow change whether we could say that the agent had in fact tried. I think this distinction fails. Consider the thoughts of the agent at the beginning of the case, having packed both the books and the laptop. Would it not be very strange to say of this agent that she did not believe, at that moment, that she would try to return the books? Surely if she did not believe that she would even try, then her decision to take the books and laptop seems rather foolish. And taking these items with her surely constitutes part of trying, regardless of whether she fails to accomplish any further actions toward returning the books. Additionally, would it not seem reasonable that the agent, if she failed to return the books, might exclaim in frustration, “I tried so hard not to forget about the damn books!” It would be strange if she could claim that she tried to overcome her predicted forgetfulness, without having a belief that she would try to do so. For this reason, I find this case, as with the previous two of the will writer and the copy-maker, fully explicable once the belief necessary for intention or intentional activity is understood as a belief to try.<sup>32</sup>

---

<sup>32</sup> A different potential strategy that could work in the favour of knowledge-backed accounts is that of Holton himself, who argues that understanding cases like the copy-maker requires us to recognise the role of “partial beliefs” in our reasoning. Holton (2009: 33) defines this notion as follows:

*Partial Belief*

One partially believes p iff one takes p as a live possibility and takes not-p as a live possibility

He further contends that it is also possible to have partial intentions, understood as intentions that are held as one amongst a number of alternatives within a given plan of action. However, not all partial intentions entail partial belief, and not all examples of partial belief give rise to only partial intention. For example, Holton takes the case of the copy-maker as one where the agent possesses an intention (non-partial given that there are no alternative intentions) based on a partial belief – the belief that he will actually make ten copies. He holds this as a live possibility, though he holds his failure as a live

A more serious potential worry, to my mind, is a point raised by Davidson (2001: 93-95) when he contends the reason that claims of ellipsis cannot be used to defend the general thesis of intention (or intentional action) as entailing beliefs to try is because not all statements of intention or descriptions of intentional action can be accurately shown to have an elliptical form. In his words:

The thesis that intending implies believing is sometimes defended by claiming that expressions of intention are generally incomplete or elliptical. Thus the man writing his will should be described as intending to try to secure the welfare of his children, not as intending to secure it, and the man with the carbon paper is merely intending to try to produce his copies. The phrases sound wrong: we should be much more apt to say he is trying, and intends to do it. But where the action is entirely in the future, we do sometimes allow that we intend to try, and we see this as more accurate than the bald statement of intention when the outcome is sufficiently in doubt. Nevertheless, I do not think the claim of ellipsis can be used to defend the general thesis. (Davidson, 2001: 92)

He proceeds to take it for granted that formulating an intention-unique belief as a belief to try is equivalent to formulating an intention with *conditionals*. By this he means something like, “I intend to go to the party if the police do not arrest me (as I suspect they may).” Further, he argues that if all intentions would be formulated most accurately by clarifying *all* conceivable conditionals, then we would be left with the “nearly empty, ‘I intend to do it if nothing prevents me, if I don’t change my mind, if nothing untoward happens.’ This tells us almost nothing about what the agent believes about the future, or what he will in fact do” (Davidson, 2001: 94). For Davidson, this

---

possibility as well. For Holton, whose interest is not in providing an account of intentional action but rather intention, this kind of partial belief is sufficient for meeting the epistemic conditions on intention. So, the copy-maker can be said to have the intention of making ten copies without violating the consistency of his intention with his beliefs. However, a proponent of a knowledge-backed account of intentional activity might go further, and argue that the presence of *at least* such a partial belief is the characteristic feature of intentional action.

There is a clear similarity between Holton’s approach and my own. Though his account is aimed at intention, rather than intentional action, we both recognise that the role of uncertainty requires that the belief necessary for intentionality (either for intention or intentional action) cannot be a full-blown belief that the agent *will do A*. The difference is that whereas Holton requires only that the agent must believe that A is a live possibility, I require that the agent must believe that they will try to A. My reason for preferring my own approach is that I take the belief to try better reflects the nature of having an intention as a commitment, as an aim, then does Holton’s partial belief. This said, I take many of the arguments I make here to be equally sustainable with the adoption of Holton’s partial beliefs, and view his approach as a live possibility in the search for an account of intentional action.

is evidence that an accurate description of an intention does not require an explication of all possible conditionals, and that the explication of a conditional should be limited to cases where the agent considers the conditional as meaningful to the pursuit of the intention, e.g. “I intend to leave the party if the music is too loud” (2001: 95). It is worth noting that this argument would also be applicable to beliefs that it is possible to A, as such beliefs seem no more or less elliptical than beliefs to try.

I think that Davidson somewhat misses the point here. Unless he is specifically targeting the views of cognitivists, where intentions merely are beliefs (which he does not explicitly claim to be doing), it is not clear how the most accurate description of an agent’s belief regarding an intention has any bearing on the resolve toward the outcome of the intention on the part of the agent. An agent can believe that she is very unlikely to succeed in A’ing, but be as resolved to it as anybody. Contra Davidson, what those defending the belief to try understanding of the intention-unique belief in the manner I am pursuing require is not that the agent’s actual intention *be* an intention to try, the intention itself can be full-blown. What is required is that the *associated belief* must be a belief that the agent will *at least* try to A, where A is the aim of the intention. It should also be stressed that the belief is a belief to *at least* try, meaning that some agents might well have stronger beliefs regarding their intentions. But such beliefs are not necessary for positive intentional status.

Another response to Davidson is to point out that he is correct: surely, it *is* a more accurate description of the intention to say, “I intend to do it if nothing prevents me, if I don’t change my mind, if nothing untoward happens.” However, Davidson’s criticism has missed the point, since what is partly characteristic about intention is that they involve – even if they cannot be reduced to – trying. In this regard, I agree fully with Brian O’Shaughnessy (1973: 365) when he stresses that “trying is an essential constituent of intentional action as such” since “no event, including intended act-events, can be foretold as an absolute certainty”. As he points out, the oddity of saying of a “normal able-bodied man in a setting of rural peace” that “they tried to walk across the road” is analogous to the oddity of saying of the President that he is sober this morning. We should see intentional action as coming with the notion of trying in-built. Note, however, that this does not mean that an intending to A is

merely intending to try to A. Rather, *part* of what it means for an agent to have an intention to A is that the agent will try to A.

Furthermore, it is not “near empty” to assert “I intend to go to the party” does not mean I believe *I will go*, or *I am certain that I will go*, but rather that I believe that *I will try to go*. Indeed, this formulation reflects the agent’s commitment to achieving the relevant intention, while allowing us to maintain the role of uncertainty in the agent’s ability to predict future action (which is only accurate to reality), while not requiring what will always be an infinite list of conditionals (which would render the statement empty), and retaining the role of commitment to action.<sup>33</sup> To see this more clearly, consider that an even more accurate description of the intention-unique belief than the one put forward by Davidson would be something like: I believe I will do it if nothing prevents me, if I don’t change my mind, if nothing untoward happens, *but I will try*.

So, though I agree with Davidson when he says of statements like, “I intend to try” that the “phrases sound wrong” (Davidson, 2001: 92), in my view, this is due to a different reason than the one proposed by Davidson. A word can feel “wrong” in a sentence for many reasons, but one of the most common is if that term makes some part of the statement tautological. I suggest that this is the case with “trying” and “intention.” For example, it certainly sounds wrong to say, “I am an unmarried bachelor.” This is not evidence that the two terms (“unmarried” and “bachelor”) in some way defeat each other, or one renders the other meaningless. Rather, it is evidence that one of the terms already incorporates the other in its meaning. If we take the intention-unique belief related to an intention or an intentional activity to be a belief to try, then a statement like:

“I have an intention to try to go to the party,”

can be taken as entailing that the agent could also make the truthful statement:

---

<sup>33</sup> I think there is also good evidence in ordinary language usage for this point. When I fail to perform an intention, it is usual to say, “I tried” as a way of expressing that I did not revise my intention, but was thwarted in what I tried to accomplish. I think it is fairly rare for a person when asked, “Why didn’t you do it?”, to respond by listing conditionals and then declaring that since these conditionals occurred the commitment to act ceased to apply. Rather, the listed conditionals may serve to indicate *why* the agent was *thwarted* (the world did not co-operate sufficiently).

“I believe that I will at least try to go to the party.”

To clarify the source of the wrongness here, let us substitute in the agent’s intention-unique belief in place of her intention in the original statement. The resulting statement reads:

“I have a [belief that I will least *try to*] *try to* go to the party.”

The second “*try to*” is redundant. It is this redundancy, I think, that lies behind Davidson’s feeling that the statement “sounds wrong.”

But, it could be asked of my solution, what about cases where we do express “try” in our statements of intention? Surely then these should also sound wrong, and the fact that they do not means that my explanation is flawed. I disagree. There are cases in ordinary language usage where the use of a redundant term serves the purpose of emphasising an aspect of the agent’s meaning, as Grice made clear in *Logic and Conversation* (1975). For example: I explain my plans to break into a house and steal some valuables to a novice accomplice. I am dubious about my accomplice’s abilities of stealth and his moral readiness to break the law, so I express part of my plan as follows: “I am going to *stealthily* sneak into the house and then *illegally* steal the jewellery off the dresser.” Now this expression clearly contains redundancies, but these are purposeful. They serve to emphasise aspects of some of the terms employed – in this case, “sneak” and “steal.” Grice (1975: 52) makes the same point when he states that although “patent tautologies” like “*Women are women* and *War is war*,” are “totally noninformative” on the level of what is said, “[t]hey are, of course informative at the level of what is implicated, and the hearer’s identification of their informative content at this level is dependent on his ability to explain the speaker’s selection of this PARTICULAR patent tautology.” In the same way, we sometimes use “trying” when seeking to emphasise the uncertainty inherent in our intention. This uncertainty is *always* present, as is the awareness of trying, but we usually only draw attention to it in the “redundant” way I have just discussed in cases of sufficient

doubt. So, it is only when I seek to emphasise to my listener the uncertainty normally implicit in my intention that I will express this intention in the form “I intend to try.”<sup>34</sup>

To summarise the takeaway point of this discussion: the belief necessarily related to any intention or intentional activity can at most be required to be a belief to try – it cannot be a belief to do. Furthermore, any account of intentional activity must be compatible with this insight. This leads me to a more complete formulation of System 2 Oversight:

**System 2 Oversight\***: A given *action X*, undertaken at time t, is under S2O, and so intentional, iff:

- (1) At time t the agent X'd
- (2) (a) X is under the guidance of the operation of System 2 or (b) is under the guidance of System 1 and System 2 had oversight
- (3) The agent's System 2 operation in this case is at least minimally responsive
- (4) The agent has a belief that she will at least try to X

---

<sup>34</sup> Grice, for his part, may have resisted my position here. In *Intention and uncertainty* (1971) he argued that a “strict intention” to A requires that the agent be sure that they will A. He bases this claim on conversations of the following type, which Davidson also employs in his arguments:

X. I intend to go to that concert on Tuesday.  
Y. You will enjoy that.  
X. I may not be there.  
Y. I am afraid I don't understand.  
X. The police are going to ask me some awkward questions on Tuesday afternoon, and I may be in prison by Tuesday evening.  
Y. Then you should have said to begin with, ‘I intend to go to the concert if I am not in prison,’ or … ‘I should probably be going,’ or ‘I aim to go,’ or ‘I intend to go if I can.’ (1971: 264-265)

Grice takes Y's remarks to be reasonable, and argues that the epistemic condition on intention should be that the agent is “sure that he will do A” (*ibid.*: 266) on this basis. But as with Davidson, it seems that Grice here is confusing an agent's commitment to do A, and her belief that she will do A. X may have every intention of making it to the concert, and yet still recognize that this could fail to occur. Given this recognition, it is reasonable for X to be less than sure that he will be there. But, given this lack of surety, we can determine that he does still intend to go to the concert if he was to say something like, “I will try to make it!” However, Y's response may be reasonable if the intention's chance of success is below a certain degree. Such cases are precisely those where we tend to explicate what is usually unspoken, that having an intention to A does not mean that I will successfully A, or that I must believe that I will successfully A (though the agent may well have such a belief), but rather than the agent is *committed* to A-ing, or *resolved* to A'ing. The belief necessarily associated with this commitment or resolve is *at least* a belief to try.

Despite the necessary role played by the belief to try in my account, I do not take it to be a knowledge-backed one. In order to see why, recall that my means of distinguishing between the various approaches discussed in Chapter 1 was by identifying what they took as the *characteristic feature* of intentional action. Further, I pointed out that few versions of any of these approaches outright deny that the features that the others hold as characteristic plays an important role. So, to discern what kind of account a given example should count as, we ask three questions: (Q1) is the feature identified as characteristic necessary for intentional action? (Q2) If both features are necessary, does one or the other have priority? And finally, (Q3) is one or the other sufficient for intentional action?

My account is a control-backed account. By this I mean that it is the presence of control that is characteristic of intentional action rather than the presence of reasons, intentions, or beliefs, though those do play a role. Given the arguments that have been discussed up till now, I will put aside talk of reasons and intentions for the moment, as they are not compelling as necessary features – we have seen that intentional action can be done for no particular reason, and with no related intention. Belief is more difficult to dismiss. Turning to the three questions, I will address them slightly out of order, going Q1, Q3, and then Q2. Is belief necessary for intentional action? Yes. Is belief sufficient for intentional action? Here I have argued no, given that it is seemingly possible to have an intention-unique belief without ever intentionally acting. I could have the belief that I will A, but in fact not A, or I could A but not *in virtue of* the belief (to use Heuer's words).<sup>35</sup> Something must be added to the mere having of the belief and the occurrence of the action in order for an intentional action to be present. This seems to indicate that the characteristic and necessary feature of intentional action is not the belief, but rather this additional feature. By contrast, intention-unique control understood as System 2 Oversight – which includes belief – is sufficient for intentional action. The fact that this control involves belief does not cede centrality to the latter, as it involves other features as well. It is this unity that is both necessary and sufficient for intentional action. This final part of the answer to Q3 indicates the answer to Q2: the reason we care about agents' beliefs regarding their

---

<sup>35</sup> This still leaves the possibility open that an intention-unique belief is sufficient for *intention*. I leave that matter to the side here, as nothing in my account hinges on whether this is so.

actions, is because it is indicative of the presence of control (or lack thereof), since the belief is a component of such control.

An attentive reader will likely at this point inquire about how my account fares when tested in this fashion against Setiya's knowing-how version of the knowledge-backed approach, given that I named this approach my chief competitor. Though I do take his account to be an undoubted improvement over previous versions, it does not threaten the centrality of control: an agent can still possess the know-how to A without intentionally A'ing, and it is not altogether clear if knowing-how can capture the cases that Setiya would want it to. Recall that knowing-how is to be understood in terms of dispositions, and so problematically limits the kinds of actions that we can call intentional. We act intentionally, at least sometimes, even when we do not have a disposition toward succeeding in the action. There may be other ways to spell out what is meant by knowing-how, but it strikes me that any attempt to provide such an explication that can allow knowing-how to fulfil the role it needs to play moves perilously close to clear talk of control. However, whether or not this is the case remains to be seen,

As it stands, System 2 Oversight\* remains a somewhat incomplete account of intentional behaviour. To see this, we can recall the cases at the fringe discussed in Chapter 1 to see if the account yields plausible answers. Though it can deal with automatic and Expressive Action, this is not the case when addressing cases of lucky action. In these cases, S2O\* seems to run the risk of overgeneration. It would *appear* to lead us to conclude that Brandon's making the shot was intentional, Nick's inputting of the correct code to thwart the meltdown would be intentional, and Aaron's killing of his target would be intentional as well. Though deeming Nick's thwarting of the meltdown as intentional might be an acceptable answer, the same is not true in the case of Brandon and Aaron. However, I take this to be only the appearance of a problem. My argument for why this criticism is misplaced begins with the acknowledgement of an important set of distinctions that has passed under the radar until this point, namely: the distinction between intentional actions, intentional omissions, intentional effects, and intentional consequences. Roughly put (to be refined below), in the cases of Brandon making the shot, Nick thwarting the meltdown, and Aaron killing the target, these are *consequences* of the agent's action

in each case. They are *not* under the direct guidance of System 2 Oversight. In these cases, the guidance exercised over the outcome is *indirect*, in that the agent did not have the ability to directly guide the outcome, but rather to guide some action that then brought about the outcome in question. So, rather than overgeneration, it might then seem that S2O\* rules out intentionality in these three cases. But this would again be missing the point: the conditions for positive intentional status though sharing certain essential elements, are not identical between the different types of intentional behaviour. What it takes for an action to be intentional is different to what it takes for a consequence to be intentional, precisely because of the different way in which the agent has control over each. Thus, in order to provide a full account of intentional behaviour, it is necessary to unpack these distinctions in intentional behaviour. Developing on this is the aim of the next two sections.

#### *1.4. Intentional activity and the stream of behaviour*

It seems to me that an important step toward a convincing account of intentional action is the realisation that what we do intentionally is not in fact limited only to *actions*. There are many things that we do intentionally. We intentionally act, as I am doing as I type this sentence. We intentionally omit to act, as when I consider whether or not to water my plant and do not do so. And we bring about intentional effects and consequences, as when I build stamina by strengthening my cardiovascular system in preparation for a marathon, or stop a nuclear meltdown by inputting the correct code. These are examples of what I will call an *intentional action*, an *intentional omission*, an *intentional effect*, and an *intentional consequence* respectively. In the language I have been using, each has a positive intentional status.

The first of these, intentional action, has – as the foregoing discussion in this chapter and Chapter 1 indicate – usually, and understandably, been taken to be the paradigmatic case of intentional behaviour. For this reason, a reader familiar with even some of the expansive literature surrounding the distinctions between so-called “mere events”, “mere behaviour”, and “intentional actions” (Ford, 2011: 76) may have been surprised at my choice to use the term intentional *activity*, rather than intentional *action*, in the title of this chapter. It should be made clear that this choice is not an accident, but reflects an important feature of my account. Insofar as we are

interested in developing an account of the unity of what is *intentionally* done, and its relationship to moral responsibility, I take it to be important that we should not turn a blind eye to omissions, and consequences. After all, it is often in the face of blame for omissions or consequences that an agent might profess in their defence, “but it wasn’t intentional!” It certainly seems that the use of the term “intentional” is expressing the same concept whether it is applied to an action, an omission or a consequence. If our interest in intentional action stems from an interest in understanding that which is “the fruit and flower of the human will” (Ford, 2011: 76), then there seems to be no good reason for excluding these other sorts of intentional behaviour. If it proves possible, all else equal, to provide an account of “doing intentionally” that retains this conceptual unity across all four these types of intentional behaviour, I take this to be a merit of the account.

What a reader will quickly realises as I begin differentiating between actions, omissions, effects, and consequences, is that these have all be readily conflated under the banner of action – excepting perhaps omission – throughout the discussion thus far. That is because, for the most part, this is the way that intentional behaviour is discussed in the literature. There are exceptions, however, notably Setiya (2012: 287) who makes a point of differentiating between the conditions necessary for positive intentional status in *basic action* and those necessary for action understood more widely. Basic action is, of course, a familiar notion in discussions of actions. By such action is usually meant *what the agent is doing not by doing something else*. So: in a situation where Nick did in fact know the codes to avert the nuclear meltdown we could construct the following chain:

“Nick averted the meltdown”  
by  
“inputting the correct code”  
by  
“the bodily movements of his fingers”

According to the conventional picture, these are not three different actions, but three descriptions of the same action. In this chain, the bodily movement of Nick’s fingers would be the basic action, as he is not doing it by doing something else. Whether all

actions can be reduced to a basic action, and whether basic action is always identified with bodily movement in this way are both open debates. While I will not be diving into these debates, it is necessary for me to be clear on how my account is be positioned relative to them. I am sympathetic to Sartorio (2007: 749) when she states:

It is natural to draw a distinction between the acts that an agent performs (his actions or omissions) and how things turn out in the world as a result of those acts (the events and states of affairs in the world – including, in some cases, acts by other people). Let's call this last category of things outcomes.

Following this, I will be referring to actions and omissions as *activities*, and to effects and consequences as *outcomes*. It is possible to translate this into talk of basic action: in which case a basic action is to be identified with a particular action or omission, and all the ‘actions’ done by doing this activity are non-basic and are all outcomes of this basic action – either effects or consequences. What marks out an activity from its outcomes is the degree of control that the agent has over it. To understand how this works, and what impact this has for understanding the control condition on positive intentional status, it is first necessary to introduce and expand on my use of the notion of *activity*.

First off, like Hornsby (2012: 234), I take it that “the agent is given her due only when it is acknowledged that she engages in activity, where no activity is any particular” and that this move to talk of activity rather than action is an ontological – in addition to a conceptual – one. Further than Hornsby, I take it to be not only that agents engage in activities, but that individual and discrete human actions and omissions are abstractions from out of the underlying ontological reality which is the continuous stream of activity. This is also akin to Stout’s (2005: 137) notion that an action may not be an “identifiable particular” but when we use the term action we are referring to “processes of acting”. The story of actions is not best understood as a man continuously walking up to crossroad and then deciding to go down a certain path, but it is rather like a man at the rudder of a ship pulled along by a current, at times sedate

and at times tumultuous, but almost never becalmed. Taking one's hands off the rudder is not a neutral or inert option, as doing so does not cease the flow of activity.<sup>36</sup>

Despite my own approach's similarity to these accounts, mine does differ from theirs in that I take omissions to be part of activity as well. So, what do I mean by activity exactly? Using this term, I seek to capture the idea that human life is *not* best characterised as a series of largely atomic actions and omissions taken against a default condition of inertness. Actual human life<sup>37</sup> is a (near) continuous stream of behaviour (which I call here activity) – which can hardly ever be fully described as “at rest” – from out of which we might abstract intentional actions and omissions. For example, “standing around” is an intentional part of the stream of activity insofar as this standing around is under the agent's control in the right way, as something that the *agent* is doing rather than doing anything else (note: *rather than* doing something else, not *by* something else – this is not a matter of identifying basic actions). On my view when we use the term “action” we are referring to an activity-slice, and the same with the term “omission.” Outcomes are somewhat different as they are indirect parts of the agent's stream of activity, the former being logical consequents of some activity, the latter being causally or explanatorily<sup>38</sup> dependent on the same.

By stipulation then, throughout my work here I will be using the term “intentional activity” to capture all instances of intentional activity-slices, i.e. intentional actions, omissions, as well as the activity taken as a whole. I will further argue that intentional effect and intentional consequences inherit their intentional status from their relationship to such activity.<sup>39</sup> I will now tackle each of these in turn, beginning with actions and omissions, then effects, and lastly consequences.

---

<sup>36</sup> I owe this metaphor to John Haldane.

<sup>37</sup> I take the points made here to also be true of non-human animal life, and possibly also the point I make hereafter about intentions often serving as a guide toward fitting action individuation, though this may be seen as more controversial as it involves the attribution of certain mental states to animals that some argue they cannot have (see Fodor, 1975 and Davidson, 1982), and others argue they can (see Bryne, Sanz, and Morgan, 2013; Plotnik, de Waal, and Reiss, 2006; Killen and de Waal, 2000). Though the question of the intentional status of the behaviour of non-human animals is undoubtedly an important one, it is a topic that I will not be exploring in this dissertation.

<sup>38</sup> I am purposefully neutral here between a causalist or anticausalist interpretation of this relationship.

<sup>39</sup> Velleman (2014) also makes a distinction between activity and action, but for him this distinction is between purposive behaviour caused by an agent's beliefs and desires, and true action, which involves causation by reasons. I take this to be a matter of stipulation, rather than revealing some meaningful disagreement between our accounts. Given his descriptions of each of these forms of behaviour, on my

In most cases when asked the question, “what are you doing?” or even, “what was she doing?” we ably pick out that activity that provides a fitting answer to the question. This is illustrated in the following passage from Anscombe (1963: 8) in which she is making a point about humans being apt at identifying the description under which an activity may be intentional:

I am referring to the sort of things you would say in a law court [if] you were a witness and were asked what a man was doing when you saw him. That is to say, in a very large number of cases, your selection from the immense variety of true statements about him which you might make would coincide with what he could say he was doing, perhaps even without reflection, certainly without adverting to observation. I am sitting in a chair writing, and anyone grown to the age of reason in the same world would know this as soon as he saw me, and in general it would be his first account of what I was doing.

We do this with amazing alacrity, regardless of whether the activity in question is extended in time, or how many steps it may involve, or whether the activity is stop-start or continuous. Indeed, two observers may well agree on what the intentional action taking place is, but strongly disagree about how to individuate the different steps involved. How and why we should be so adept at this is a matter deserving its own discussion. But what is pertinent for my arguments here is that, in many cases, under one of the available descriptions the action in question is an intentional action. Indeed, this will often (though by no means always) be the “first account” which we strike on. What is more, and leaning again on Anscombe’s insights, is that the presence of the positive intentional status guides the individuation process. I discern that when asked in a normal way, “what is your friend doing?” the fitting way to describe the certain slice of activity my friend is engaged in is with the description, “he is making tea,” rather than, “he is shunting air particles,” precisely because, seemingly, it is the agent’s intention to do the former, not the latter. Simply put, we cut some actions (as well as omissions) out of the stream of activity through the guidance of the possession of positive intentional status. Actions or omissions

---

account they would translate to unintentional activity and intentional activity respectively. A reader should feel free to plug in whichever terminology she or he prefers.

individuated in this manner are, if we are accurate, intentional action or intentional omissions. Steward (2012: 384) makes this same point, when she says:

Just as, in thinking of a horse, say, we lock onto a creature whose principle of individuation has to do with continued life and not with continuation of the very same matter, so in thinking of a process, we lock onto an entity which we conceive of as having a principle of individuation which has to do with what one might call norms of development – or, in the case specifically of human action, with such things as intentions and goals.

On my account activities are processes, not events, and these are singled out by some guidance process which is directed by certain normative expectations – what Steward calls “norms of development”. This gives us a more robust way of speaking about actions that are, after all, often extended and ongoing in time and open to in medias res adverbial modification (i.e. the action can be performed faster, or slower, more vigorously or less), both aspects that the actions as events view struggles to account for.<sup>40</sup>

Following from this, and my motivation not to exclude intentional omissions from the picture I am sketching, I take an intentional omission to be an abstraction from the stream of activity in a manner not relevantly different from an intentional action, and that like such actions, such individuation can be – indeed usually is – guided by the (supposed) presence of positive intentional status. One consequence of this is that I argue that we can distinguish a set of omissions as intentional omission and that this set has meaningful content.

It is important to note that I am not claiming that all cases of action or omission individuation are guided by intentional status, nor that such guidance is always the best route to take. After all, we are fully capable of identifying *unintentional* actions, omissions or mere behaviours. There are any number of different ways to individuate actions and omissions, as Anscombe alludes to in the extract. To build on her point, the question: “What was the man doing?” can be ambiguous outside of the courtroom

---

<sup>40</sup> See Steward (2012) and Hornsby (2012) for good overviews of these shortcomings in the actions as events approach.

scenario she sketches (though even in a courtroom it may not always be unequivocal just what the question is aimed at). In one sense of what the man is doing he will likely be shunting air particles. Or say that the man in question unknowingly, and unintentionally, steps on another person's foot. In this case, one answer to "what is the man doing" would be "stepping on the other's foot", and this is picked out regardless of intentional status. To make this last point even sharper, consider that we could individuate out and attribute the *behaviour* of "stepping on the other's foot" to the agent even if it resulted from someone else pushing him, or him slipping on a slippery surface – in other words, it was not the result of "the fruit and flower of the human will". What is true is that, in this latter case, facts about the origination of the behaviour – which is partly constitutive of it being an example of mere behaviour – do undermine the degree to which we may take the agent to be responsible for the behaviour, as well as the degree to which the behaviour is seen to reflect the agent's character or agency. Though more pronounced in intensity, these are also common features of unintentional action. However, what matters for my discussion here is that we are (for the most part) perfectly capable of identifying the unintentional action or behaviour of stepping on the other's foot, even though intentional status plays no guiding role in the individuation in either case.

It is also the case that we can individuate out actions without intentional guidance even when the action so individuated may well be intentional when considered under such guidance. So, imagine a case where a doctor is examining an agent's lifestyle to give them advice on healthy living. During the examination, the doctor identifies that the agent often begins typing at her desk at a good distance from the screen, but then leans her head in closer to the screen during the work period – with negative results to both her posture and her eyes. This leaning in could be intentional or not, however, the doctor can individuate out, and evaluate the health consequences of, that slice of activity without the need for intentional guidance. Additionally, whether that given activity was merely behaviour, or an unintentional action, or an intentional action, will depend not on the guidance used to individuate it, but rather on facts about the activity at that moment. My claim is only that we *often do* individuate actions and omissions through the guidance of intentional status, and that *when we are accurate* in doing so, the given activity will be an intentional one.

While there are numerous ways to individuate activity-slices, in certain contexts some seem more appropriate than others: in Anscombe's example of a court of law, intentional guidance seems the most appropriate, in the case of the doctor I sketched above, we might say that the appropriate method was something like physiological guidance – where the mere physical facts about bodily movement serve to guide individuation in some way. The most appropriate method for individuation in any given scenario will depend upon various contextual features.

As should be clear at this point, I will defend the argument that some slice of activity (be it an action or omission) is intentional not because it simply follows from an intention, was done for reasons, or the agent knew that she was doing it, but because that slice of activity was *under a certain kind of control*, namely: System 2 Oversight. A given agent's activity-slice could be intentional even if the agent had no preceding intention or reason to act at all, provided that the correct sort of control was present. This is not to say that intentions or reasons to act are not important. It is true that often our intentional activities follow from our intentions, and that when we act under the (non-deviant) guidance of an intention such activity will necessarily be under System 2 Oversight – I discuss this more in Section 2.2. And as I have already shown, reasons to act do play a central and often essential role in the kind of control necessary for positive intentional status. To reflect all the above, the provisional principle IST that was introduced in Section 1 must be widened to account for the inclusion of omissions:

**Intentional Status Transmission (activity):** An agent's action or omission is intentional iff the agent possessed intention-unique control over it.

And we can go further and replace intention-unique control with System 2 Oversight, yielding:

**Intentional Status Transmission (activity<sup>\*</sup>):** An agent's action or omission is intentional iff the agent possessed System 2 Oversight over it.

IST(activity<sup>\*</sup>) is also the measure against which our ascriptions of intentional status are to be judged. When we ascribe intentional status to a given agent's action, or

omission over which the agent in question does not possess System 2 Oversight then we are in error – though this could, of course, be blameless error depending on the context. Said differently, I would be cutting up the stream of activity incorrectly if I failed to individuate intentional activities per the guidance of IST(activity\*).

In the majority of cases where an agent has System 2 Oversight over some activity A, she will also have an intention either to A, or to B where A is a means or foreseen effect of fulfilling this intention. Importantly, this does not mean that we first ascribe an intention and then set about working out what are the intentional actions, etc. Rather what happens is that when we encounter an agent amid activity, then we will usually assume (barring the agent being asleep, being carried away by an avalanche, i.e. in a state that undermines her agency and breaks the stream of activity) that some activity-slice from her stream of activity will be intentional. The assumption that positive intentional status is present is the default. From experience, we will have a fairly accurate grasp of how to slice an agent's stream of activity in order to individuate out what is intentional – or in Anscombean terminology (1963: 46-47), we are effective at identifying the description under which the correct action or omission is picked out as intentional – and so will usually be guided by this expectation of intentional status in individuating out a given slice of intentional activity. While we are doing this, we will also be ascribing an intention to the agent, one that would stand in the appropriate relationship to the activity we have individuated out as intentional. We can, of course, be mistaken, but what is crucial is that the IST(activity\*) is the bar we must measure our individuations and ascriptions against.

This also allows us to spell out a more complete formulation of System 2 Oversight:

**System 2 Oversight\*\*:** A given *action or omission* X, undertaken at time t, is under S2O, and so intentional, iff:

- (1) At time t the agent X'd
- (2) (a) X is under the guidance of the operation of System 2 or (b) is under the guidance of System 1 and System 2 had oversight
- (3) The agent's System 2 operation in this case is at least minimally reason-responsive

(4) The agent has a belief that she will at least try to X

For a given action X, the X indicates what it is the agent is doing. For a given omission, X, the X also indicates what the agent is doing, namely:  $\neg Y$  (where Y is some action). Because of this, the belief that the agent will try to X in the case of an omission would be a belief that the agent will try to  $\neg Y$ , where Y is what the agent is omitting to do. So, for example, if I decide not to water my plants, I must have the belief that I will at least try not to water my plants. Having dealt with intentional actions and omissions, I turn now to intentional effects and consequences.

### *1.5. Accounting for intentional effects and consequences*

The difference between the effects and consequences of an activity is quite often overlooked, but it is an important distinction to keep in mind. Consequences are usually what we think of when we consider the outcomes of an activity, these being the causal results of the action or omission,<sup>41</sup> or the results that are grounded by the action or omission.<sup>42</sup> On the other hand, effects are not *caused* by an action or omission, but are rather the logical consequents by said action or omission.<sup>43</sup> An example from Sartorio (2010: 576):

Imagine that an assassin murders a child's parents. When he does, the child becomes an orphan. But the assassin's murder of the parents doesn't cause the child's orphanhood, for the fact that the child's parents were murdered entails the fact that the child is an orphan, and entailment is not causation.

---

<sup>41</sup> Whether or not omissions can be causes is a very debated topic, and perhaps one of the reasons militating in favour of the anticausalist position. For my purposes, I will be assuming that either omissions can be causes, or that the relationship between activities and consequences is non-causal.

<sup>42</sup> Since I aim to be agnostic in the causalist/anticausalist debate, I will not be adjudicating between these two understandings of how consequences related to activities. From this point on I will be employing causalist language, but this is merely a decision made for ease of writing.

<sup>43</sup> Another way of understanding this distinction is outlined by Gormally (2016: 305-308) in his discussion of Anscombe's approach to the morality of killing human beings. Following Anscombe he contends that: to be doing some things is sometimes, *eo ipso*, to be doing some other thing. To crush an infant's skull *is* to kill the infant. In these cases, the agent has "certainty" about the outcome if she is aware of it. In contrast, there are other things that are done by doing other things that do not have an *eo ipso* identity in this way. If I fire an arrow from a bow, it does not certainly (I would say necessarily) follow that "a man be transfixed."

The first thing to note is that whereas intentional activities are directly controlled, intentional effects and consequences are controlled in a mediated or indirect fashion. Furthermore, an agent has more proximate control over the intentional effects of her activities than over their intentional consequences. This has important implications for thinking about the control condition applicable in each instance. I will discuss intentional effects before coming to intentional consequences, as in many ways the latter is more difficult to account for than the former.

Where effects are concerned, the most obvious difference to intentional activity is that such an effect is the *logical* consequence of some intentional activity, where this relationship is not a causal, or grounding relationship. The fact that a given intentional effect is brought about is logically entailed by the fact that a given intentional activity took place. We do not take the effects of unintentional activity to be intentional. However, the transfer of positive intentional status from activity to effect is not a necessary transfer, nor is the fact that a given effect is brought about by an intentional activity sufficient for the effect to be intentional: there are many effects of an agent's intentional activities that do not count as intentional. The reason for this is that intentional effects are not under the guidance of the System 2 Oversight mechanism in the same way as the action or omission is. Only those effects that are under the guidance of System 2 Oversight, such that they can be brought about with sufficient reliability are considered intentional. Recall: for an action or omission to be under the guidance of System 2 Oversight, it is necessary that the agent be able to adjust the action or omission in the light of reasons. Unlike actions and omissions, however, effects can only be guided by *adjustments to the actions or omissions that bring them about*. So, in order for an effect to be intentional, it must be the case that the agent had the capacity to respond to the available reasons to act such that she could adjust her action or omission in light of the effects thereof with sufficient reliability. For this to be possible, the agent must not only believe that she will try to bring about the effect, but must also have the skill and knowledge necessary to do so with a sufficient degree of reliability as to be responsive to the salient reasons.

Consider Nick at the keyboard. His inputting the correct code is a logical consequence of his striking the correct combination of keys, yet we wish to resist the idea that he intentionally input the correct code. Yet, his inputting the correct code was the result

of the System 2 oversight mechanism – in this case a direct result of the operation of System 2, since his inputting digits was presumably under conscious control – this mechanism was suitably reason-responsive, and Nick certainly believed that he would at least try to input the correct code. So, is this an intentional effect? No, because though it is the result of the operation of System 2, being the logical consequence of Nick’s intentional action of typing in digits on the keypad, *inputting the correct code* was not under guidance such that he could have brought it about reliably. Inputting the code was not under guidance for the simple fact that Nick did not know that what he was doing in inputting the ten-digit sequence was in fact inputting the correct code, and so we would not have been appropriately responsive to reasons to adjust his action at the time of typing to insure his success with sufficient likelihood. Nick could not respond to these reasons – such as, say, the third digit of the code is 4, thus there is a reason to input 4 on the third input – as without knowing that the code in question was in fact the correct one, his System 2 Oversight mechanism would not be receptive to reasons to adjust his actions in favour of it.

This gives us what I take to be the final version of System 2 Oversight, which I take to capture intentional actions, omissions, and effects:

**System 2 Oversight (complete):** an agent’s X’ing, undertaken at time t, is under S2O, and so intentional, iff:

- (1) At time t the agent X’d
- (2) X is under some degree of guidance by the operation of (a) System 2 or (b) System 1 with System 2 oversight, such that the agent could bring it about with sufficient reliability through this guidance at time t
- (3) The operation of the agent’s System 2 is at least minimally reason-responsive
- (4) The agent has a belief that she will at least try to X

From here on I will simply refer to this as System 2 Oversight (S2O). This then leaves us with intentional consequences. Continuing talk of the example of Nick, consider that just as we do not want to call his inputting of the correct code intentional, it does seem incorrect to say that he intentionally averted the meltdown. I take this averting of the meltdown not to be an intentional consequence of Nick’s intentional activity,

which was the inputting of the digits into the keyboard. It is often the case that the intentional consequence of an action or omission might be decidedly *outside* the control of the agent. Indeed, the mark of a consequence is that it usually requires the world to “play along” to a far greater extent than required by intentional activity. This is most vividly illustrated by cases where the agency of others plays a large role in determining the outcome: if I bowl a delivery in cricket, which the batsman nicks on to his own wickets, it would be said that this is an intentional consequence of my action in bowling the ball. Yet, it would not have occurred if it had not been for the batsman’s own actions. It should also be noted (as came up in the discussion of “lucky” action) that in these cases much seems to depend upon the intent or beliefs of the agent: if I bowled the ball looking to get the batsman to draw the ball on to his wickets, but he instead manages to catch the ball more fully and hoist it into the air to be caught out, it seems less obvious that we should count this as an intentional consequence. Yet certainly not only consequences that are *intended* should count as intentional. Bratman (1984) is surely correct in urging the abandonment of the assumption of tight fit. However, Bratman’s approach still leaves us with the puzzle of spelling out the criteria by which a given consequence (or indeed action) falls within a given intention’s motivational potential. And there is the further problem that – just as with intentional activities – not all intentional consequences necessarily follow from an intention.

Again, we must turn to the role of activity. The consequences of an activity (be it an action or an omission) are those states of the world that follow causally from the activity, or that are explained by them. So how we individuate actions and omissions from the stream of activity will influence what we think of as the relevant consequences. In the case of intentional consequences, whatever else they may be, they must follow appropriately from an intentional action or omission. But there must clearly be more to the story, since many of the consequences of an intentional activity are certainly not intentional: if I wave my hand to greet a friend, and this waving results in a hurricane in Sri Lanka, it would be implausible to say that the storm was an intentional consequence of my action. Indeed, it seems to be a paradigmatic case of an unintentional consequence. Yet, if we take the same example, but this time say that I am the world’s greatest meteorologist, and have discerned with absolute certainty that my waving will indeed cause such a storm, and I wave regardless, then it seems

entirely plausible that bringing about the hurricane could be described as an intentional consequence of my action. This comparison helps to highlight that what appears critical here is whether I *foresaw* the consequence in question, and that this links to my having *sufficient control* over it.

In terms of epistemic status, consequences can be divided into four overlapping categories: unforeseeable consequences, unforeseen consequences, foreseeable consequences and foreseen consequences. An unforeseeable consequence is one that it would have been impossible for the agent to foresee. An unforeseen consequence is one that the agent in fact did not foresee. A foreseeable consequence is a consequence that it is possible for the agent to foresee. And finally, a consequence is foreseen if the agent is aware of the possibility of it occurring as the result of her actions or omissions. Now clearly unforeseen and foreseen consequences are mutually exclusive, as are unforeseeable and foreseeable consequences. However, an unforeseen consequence can be either unforeseeable or foreseeable. Similarly, a foreseeable consequence can be either unforeseen or foreseen. An unforeseeable consequence is always unforeseen, and a foreseen consequence is always foreseeable. If a consequence is unforeseeable, then it falls outside of an agent's control, and it is entirely implausible to think that it is intentional. Unforeseen but foreseeable consequence are equally unintentional as they again fall outside of an agent's System 2 Oversight, though we may be inclined to misattribute positive intentional status in these cases. I will be returning to this particular type of case, the unforeseen foreseeable, in my discussion of moral responsibility in Chapter 3: Section 5.1.2., as the most compelling counterexamples to the idea that such consequences can be intentional involve instances where the consequence is morally bad in some way.

However, in the case of foreseen consequences, if the agent foresaw the consequence of her activity, provided that the activity itself was intentional, then this consequence was one over which the agent had *some degree* of discretionary control. This follows from the fact that if the consequence was foreseen, the agent would have had to have chosen to continue in undertaking the relevant activity knowing that the consequence could occur. Now it is of course the case that different consequences have different likelihoods, and that this does seem to be a relevant consideration when determining whether or not a given consequence should count as intentional. After all, I know that

there is a (minuscule) possibility that a consequence of my next keystroke might be that a hurricane in fact hits Sri Lanka, through some intricate causal chain of which the air movement of my keystroke is a tiny part, yet it would be implausible to say that this would be intentional.

Consider also the following, it is not sheer unlikelihood that undermines positive intentional status. To see that this is the case, take examples where the odds of success are equal, yet our intuitions about intentional status differ:

*Golfer*: a professional golfer, Gwen, is preparing to make a very difficult putt. As it happens, Gwen has a 1-in-6 chance of sinking the putt. She sinks the putt.

*Dicer*: a person, Dina, picks up a die as part of playing a board game. She needs to roll a six to win. Dina clearly had a 1-in-6 chance of rolling a six (if the die is perfectly balanced and Dina does not know any rolling tricks). She rolls and gets a six.

Gwen seems to sink her putt intentionally, whereas Dina does not roll the six intentionally.<sup>44</sup> But what is different between these two? In both cases the outcome is equally foreseen, and the odds are the same. The answer is the role of control. For a given consequence to be under sufficient control to be intentional, it must be the case that the agent's skills, and beliefs played a sufficient role in determining the likelihood of success. In the case of Dina, her skills, beliefs, and desires play no role in determining which side of the die ends up facing up, whereas in the case of Gwen this is not so. Had Gwen been an eight-year-old trying golf for the first time her odds of success would have been dramatically lower than 1-in-6, so her control plays a sufficient role in determining the success of her putt. As we have already discussed, the way that an agent's skill, and belief guides behaviour is through the guidance of System 2 oversight, and in the case of intentional consequences, this guidance is indirect. My skills and beliefs influence my chance of bringing about a given consequence through how receptive and reactive I am to the various reasons that would, were I to recognise and react to them, increase my chance of *acting* (or

---

<sup>44</sup> This example can be made even more vivid by replacing the die in *Dicer* with a perfectly weighted coin. Here, even though Gwen has a lower likelihood of success, it still seems that her making the put was intentional, whereas Dina getting any particular coin facing is not.

*omitting) such as to bring it about.* This control is indirect because it operates *through* my actions and omissions.

It could be objected to this that one can equally well learn how to roll dice skilfully to land on certain facings rather than others, so it isn't that rolling dice has no role for control. However, note that if Dina was such an expert at rolling, then we would indeed be more willing to attribute positive intentional status to her rolling a six. This also explains why the lottery case seems to clearly not be a case of intentionally winning – in this example the agent obviously and (if it is to be a truly fair lottery) *necessarily*, has no control over the outcome.

To capture this reality, I propose the following as a transfer principle for positive intentional status, transferring it from the intentional activity to the consequence:

**Transfer of intentional status (TIS):** a given consequence, X, of an agent's activity, Y, occurring at time t, is intentional iff:

- (1) The agent has the belief that bringing about X is, given her evidence, a *possible part of the fulfilment of an aim to Z*, which involves Y'ing
- (2) X is brought about under *sufficient* guidance of System 2 Oversight by Y'ing, in line with the agent's plan to Z

Y and Z can be the same. Before continuing it is necessary to explain the notion of *plan* as it is used here: an agent's plan involves some aim and a set of related beliefs about how to achieve the means to that aim and in turn how to achieve these means. Such plans can – and often do – involve intentions. Indeed, I am sympathetic to the view that such plans that precede intentional consequences *always* involve intentions (Mele and Moser, 1994; Bratman, 1987; Mele, 2009). However, I see no reason to marry my account to this position, and so will not take a stand on this question here. Such plans can obviously come in varying degrees of reliability. They are also the result of the operations of the System 2 oversight mechanism, and as such are reason-responsive. This responsiveness is limited by the state of the agent's beliefs. For plans like this (what I will sometimes call *action plans*, by which I do not mean that they are limited to *actions*) to be *as reliable* a plan as could be expected of the agent means

that the plan must maximize the agent's chances of fulfilling its aim given the agent's set of beliefs about means – and so the reasons that the agent could be responsive to.

The transfer captured by this principle is that of revealed rational agency. The given consequence reflects and exhibits the agent's locus of agency only when the stated conditions are met. Where an agent has an action plan to  $Z$  that involves a step,  $X$ , where the agent is incapable of exerting sufficient control, then that step – even if successful – will not count as intentional. The reason that the uncontrolled step fails to qualify as intentional is because  $X$  occurring is not sufficiently indicative of rational agency. So, when an agent enters the lottery, for example, her entering the lottery is intentional, but her choosing the winning ticket is not, since which ticket she chooses tells us nothing about the agent's rational agency. It tells us nothing in this regard because the agent could not be responsive to the reasons for and against picking any particular ticket. As goes the step, so goes the plan. If a step in an agent's plan succeeds, but this success is sufficiently divorced from the agent's control, then the success of the plan, taken as a whole, will also fail to be under sufficient control. I leave the degree of control necessary for sufficiency purposefully vague, as I believe that there is (possibly irreducible) vagueness inherent in what we take to be a sufficient degree of control. Also note, there is an interplay between the odds of success and the degree of control, but this interplay does not stand in a simple relationship with positive intentional status. The odds of my success might well be quite good (rolling higher than one on a regular die, for instance), but if there is no control exerted then we would not think of a successful result as intentional. On the other hand, the odds might be quite poor (a long difficult putt that the putter only has a 1-in-50 chance of making), but the putter has control over the outcome such that were it not for this control the odds would be a million-to-one, and so it seems plausible that we would think of the successful putt as intentional. There is undoubtedly much fertile ground to still investigate here, but I delve no deeper into it in this dissertation, but rather turn to see how the TIS deals with the lucky cases, the by now familiar examples of Brandon, Aaron, and Nick.

We can imagine a case where Brandon believes that he will make the million-to-one shot, in that case I diagnose the situation as follows: Brandon's making of the shot meets condition (1), but although it is brought about in line with his plan, he has

insufficient control over making the shot for it to count as intentional. Brandon's skills and beliefs, his receptivity and reactivity to reason, simply have too little influence on the outcome for it to be intentional. Were the odds to be reduced to a more plausible level, however, then the intentionality of making the shot would depend upon the level of control that Brandon exerts over its being brought about, relative to the odds.

Aaron believes that killing Target is a sufficiently likely part of the fulfilment of his intention to kill Target, which involves shooting him. However, his shooting of Target is neither brought about in line with his plan nor under sufficient guidance from System 2 Oversight.

Lastly, Nick will believe that inputting the correct code is a likely (indeed necessary) part of the fulfilment of his plan, which involves inputting digits into the console. However, his actual inputting of the correct code, though following in line with his plan, was not under sufficient guidance from System 2 Oversight, since he was unresponsive to the reasons salient to picking the correct code – due to the fact that he did not know what it was. At the same time, were he to succeed in inputting the correct code, his averting of the meltdown would also not be intentional, as it was brought about in line with an agent's action plan which involved a step over which the agent had insufficient control, and so it itself was not brought about under sufficient control.

Given its ability to capture plausible answers to the various cases that have been considered, I take TIS to provide a principle that exhaustively accounts for intentional consequences. Combined with S2O(complete), I take all intentional behaviour to have been accounted for.

### *1.6. Concluding remarks*

In this section I have presented my own account of intentional activity (and intentional outcomes), which I call System 2 Oversight. I have argued that this is a control-backed and mechanism focused account, where the mechanism that exerts intention-unique control is best thought of as System 2 oversight, a position built on

insights drawn from Dual Process Theory. I have further contended that this understanding of the mechanism in question meets four crucial requirements: it is suitably reason-responsive, it accounts for the epistemic condition on intentional activity, it does not prevent the possibility of an agent acting against the balance of reasons, and finally it allows that the mechanism can result in both intentional activity as well as intentions and not necessarily together. In the course of doing so, I presented arguments defending my control account against notable objections, as well as indicating some of the shortcomings of my account's competitors. Lastly, I presented my transfer principle for intentional consequences, TIS, whereby such consequences derive their positive intentional status from the relationship they must have to the agent's plans, and thus by extension the agent's intention-unique control mechanism. Part of understanding this transfer principle depends upon a certain technical understanding of the notion of *plan*, which I outlined. To adduce further evidence in support of my account, in the next section I will demonstrate how my control account manages to fulfill an additional criterion that I take to be necessary for any convincing account of intentional activity: explaining the unity of the three applications of the concept intention.

## 2. The unity of the three applications of the concept intention

Ever since the publication of Anscombe's *Intention* (1963), there has been near unanimous agreement amongst philosophers that any convincing theory of intentional activity, or of intention, must meet at least one key criterion: it must explain the unity between the three seemingly irreconcilable applications of the concept intention, namely as "intention-for-the-future, intentional action and as the intention with which someone acts" (Anscombe, 1963: 1), or present convincing reasons for forsaking this unity.<sup>45</sup> Call this the *Unity Criterion*. In this section, I attempt to meet this criterion by explaining the unity of the three applications as consisting in their shared relationship to the right sort of control, namely: System 2 Oversight.

Additionally, there remains an important potential criticism of my account that I have not yet presented a complete answer to: the problem of mutually exclusive intentions.

---

<sup>45</sup> For examples of those arguing to retain the unity, see Davidson (2001), Velleman (1989), Bratman (1985) and Setiya (2011). For examples of those arguing that we should reject the unity, see Knobe and Burra (2006), Harman (1986; 2006), and Holton (2009).

Since this problem involves the relationship between different uses of intention, I will endeavour to show that my account of the unity of the three uses can resolve it satisfactorily.

### *2.1. Introducing the three applications of the concept intention*

Right at the beginning of *Intention*, Anscombe observes that there are three common ways in which we employ the concept intention:

1. The agent intends to G (verb)
2. The agent G'd intentionally (adverb)
3. The agent F'd with the intention of Ging (noun)

The first use of the concept (1), which Anscombe defines as intention-for-the-future or prospective intention, refers to the fact that we commonly say “I intend to do X at time Y,” where Y is still to come. I can have such an intention for a very long time before I take any actions to bring about X; for example: as a sixteen year old I may intend to become president one day, and even though I take no immediate action towards that goal until I am considerably older, it remains the case that I intend to do so (become president). Conversely, prospective intentions could be only very slightly anterior to the performance of action, such as forming the intention to scratch my back and doing so (almost) immediately. As will become clear, there is much contestation about exactly *when* such an intention can be said to be present, and whether the presence of such intention requires that some action takes place at some point. For example, if I never actually take steps to become president, can I ever be said to have had the intention of doing so? It seems obvious that the intention entails a sense of commitment, but must that commitment manifest in action for it to qualify as such?

The second use (2) refers to intentional action, meaning that we identify a given action as taking place due to *human agency of a certain sort*, as opposed to by accident or due to forces of nature external to the agent. For example: I lift my hand intentionally as opposed to it being lifted by somebody else, or it being lifted due to an involuntary reflex. It has been this use that I have focussed on, and for the reasons

outlined in Section 1.4. I will be speaking of intentional activity, rather than intentional action to refer to this use from here on. It is this use of intention that is most frequently rejected from the unity. That is, several thinkers argue that whether or not an activity is intentional is not necessarily conceptually related to whether the agent had an intention.<sup>46</sup> I take the question of what exactly is required for some activity to count as intentional – that is, what is meant by *human agency of a certain sort* – to be the most important question that an account of intentional activity must answer, and I take the answer to be the presence of a certain kind of control: System 2 Oversight.

The third use (3) is intention-with-which, i.e. the intention I have when performing an action. This use of intention is often used in a teleological sense; that is, it is used to describe an action as directed toward the achievement of a certain goal. For example: I am reading an article by Donald Davidson because I want to better structure my analysis of his arguments.

The problem is that these three uses seem to be “not equivocal,” and our inability to explain the rules underlying the use of the term “intention” in these three different ways reflects that “we are pretty much in the dark about the character of the concept which it represents” (Anscombe, 1963: 1). Until these underlying rules are understood – or until we are presented with good reasons to dismiss the intuition that they are related – a holistic theory of the concept intention that can convincingly account for all three uses is unachievable. For this reason, addressing the disparities in our use of the term remains one of the fundamental requirements of any convincing theory of any one of the uses of the concept intention. And thus, as my aim is to present a

---

<sup>46</sup> Much of the support for this view in recent times has stemmed from a series of experiments conducted by Joshua Knobe (2003; 2004) together with Arudra Burra (2006), which are commonly interpreted as showing that folk attributions of “intention” and “intentionally” come apart. I will not be presenting my arguments against Knobe and Burra’s conclusions here. However, I will briefly state that I do not take their experimental results to unequivocally support the conclusions that are derived from them, and that by adopting a wide-fit view of the relationship between intentions and intentional actions it is possible to account for the discrepancy that they identify in the attributions. Furthermore, their argument that the moral valence of an action influences attributions of positive intentional status to actions, but not attributions of intention, can be explained by the “pragmatic considerations” (Adams and Steadman 2004 and Adams 2006) – indeed, I take Jennifer Nado to present a compelling account of how these considerations might “unduly” influence our attributions of positive intentional status in her 2008 work *Effects of Moral Cognition on Judgements of Intentionality*.

convincing account of intentional activity, it behoves me to lay-out my proposal for such a unity.

However, to complicate matters even further, Wilson and Shpall (2012) contend that there is an additional use of intention for which Anscombe had failed to account. They state this use to be the following: (1) “in *F*ing (by *F*ing), the agent intended to *G*”, which they argue is related to but distinct from intention-with-which. Given that the most basic expression of intention-with-which is: “the agent *Fed* with the intention of *Ging*,” they employ the following example to illustrate their point:

[A]lthough it may be true that

(8) Veronica mopped the kitchen then with the intention of feeding her flamingo afterwards,

it normally won't be true that

(8') In (by) mopping the kitchen, Veronica intended to feed her flamingo afterwards.

The irreconcilable nature of these two statements is meant to show that the latter use of intention must be considered distinct from the first, and so should be added to Anscombe's list of uses. However, this would be a mistake. In both examples the two intentions – mopping the floor and feeding the flamingo – are separate intentions, and so what is at stake here is not some new use that seems close to a case of intention-with-which, but rather a statement that includes an intention-for-the-future and an intentional action. The action of mopping the floor does not bring about or aid the progress of the action of feeding the flamingo, or if it does then the statement (8') would not be problematic in the sense Wilson and Shpall take it to be. The problem centres around the word “with” as it is used in the first statement, which seems to indicate a relation to intention-with-which, but this misses the fact that there is a difference between doing an action *with* an intention in the sense of “I have another intention while I am separately and intentionally performing my current action” and doing an action *with* an intention in the sense that “I am performing my current action with the intention of performing another.” The first type does not imply a connection between the two actions, the latter does. Assume a case where the flamingo cannot be fed until the floor is clean. In that case, it would be perfectly correct to say that, “in

mopping the floor, Veronica intended to feed her flamingos afterwards,” precisely because here the two actions are linked, and the intention to feed the flamingos is the intention *with which* Veronica mops the floor. Let us call this case (8’’). Assuming, as Wilson and Shpall do, that no such necessary connection as in (8’’) exists between Veronica’s actions, we are left with not one new use of intention, but a statement that expresses two different applications of the concept intention, each application relating to a given intention. The mopping of the floor is a case of intentional action: “I am (intentionally) doing X,” and the feeding of the flamingo is a case of intention-for-the-future: “I intend to do X.” In the case of (8’), intention-for-the-future has simply been given in a reported form: “she intends to do X.”

What this analysis shows us is twofold. First, it reinforces Anscombe’s position concerning the three – and only three – applications of the concept intention. Second, and more importantly, it raises the fact that we can, and commonly do, hold multiple intentions. Statements such as (8’) are complexes in the sense that they refer to multiple intentions. The statement as a whole is not an example of any *one* use of the concept of intention, but each intention in the statement is itself an example of one use of the concept. This is a concern that Anscombe herself does not address directly. It is likely that there are rules governing the relationships between multiple intentions (such as are present in (8) and (8’)) that we ought to try to explicate. This question is best addressed not under the criterion of the unity of the three uses, but as part of a consideration of the relationship between intention and practical reason. I will not be investigating this question here, and at this point, it is enough merely to note the existence of these complexes, as this supports the view that there are only three applications of the concept intention.

## *2.2. Unity through control*

I start my attempt at developing a unity of the three applications of the concept intention that can satisfy the Unity Criterion, by rejecting the idea that the unity is to be found by the reduction of the three uses to one use. It is not the case that every intention-for-the-future requires an intention-with-which or an intentional action, nor that every intentional action requires an intention-for-the-future. That said, I will argue that intention-with-which does stand in such a necessary relationship, as this

use of intention amounts to the explanation of an intentional action in terms of its related intention-for-the-future. Overall though, rather than a reduction to a single use, my argument is that the shared notion of “intention” at work in all three these uses is best explained in terms of each use’s relationship to System 2 Oversight. Given that I have already unpacked this relationship for intentional activity (which included intentional action), I will proceed by first discussing intention-for-the-future, and its relationship to System 2 Oversight and to intentional activity, and then move on to intention-with-which.

In our daily life, intentions-for-the-future (from here on simply *intentions*) are an indispensable tool for achieving our practical goals, particularly as regards cross-temporal planning. When I describe my commitment to performing an action it is commonly in the form, “I intend to do it.” My plans for the future invariably take the form of a series of intentions, which though open to revision, guide actions. Whatever else it may be, intention seems to clearly be a practical or optative attitude, one that is inextricably linked to my actions and to my practical reasoning. Given this reality, I take an intention to be a unique optative mental state built up of a conduct-controlling pro-attitude and a related belief, where the belief can be described as: the belief that I should at least *try* to fulfil [what the relevant pro-attitude aims at]. What marks this mental state out as an *intention* is its relationship to the agent’s rational agency, which is to say: the mental state is the result of a System 2 oversight mechanism. Furthermore, intentions are, as Bratman (2009a) has argued, our means of rationally controlling our conduct in so far as they are open to certain rational requirements (or regularities). This implies that what separates an intention from other pro-attitudes is not, as it is for Davidson (2001), that it is identified with an all-things-considered judgement, nor as Velleman (1989) thinks, that the intention be identified with a particular belief. Rather, what is characteristic about those conduct-controlling pro-attitudes we call intentions is that the agent possesses a certain type of control over it, and that they in turn play a crucial role in controlling and guiding our conduct.

Of these rational requirements, I will only address what I take to be the most important two, both of which have already been introduced briefly in foregoing sections. The formulation of the first requirement presented here is taken from Broome (2013: 159-170), while the second is from Bratman (2009b: 413):

*Instrumental rationality:* Rationality requires of  $N$  that, if,

- (1)  $N$  intends at  $t$  that  $e$ , and if
- (2)  $N$  believes at  $t$  that, if  $m$  were not so, because of that  $e$  would not be so, and if
- (3)  $N$  believes at  $t$  that, if she herself were not then to intend  $m$ , because of that  $m$  would not be so, then
- (4)  $N$  intends at  $t$  that  $m$ .

*Intention Consistency:* The following is always pro tanto irrational: intending A and intending B, while believing that A and B are not copossible.<sup>47</sup>

As with intentional activity, the kind of control we have over our intentions is not always conscious control. There are undoubtedly many cases where an agent might form an intention without being consciously aware at that time of having done so. That said, I do take it to be the case that any intention must have the possibility of being brought to consciousness. As with intentional activity, I take intentions to be one of the possible outputs of an agent's dual cognitive systems, System 1 and System 2. Also in line with my arguments pertaining to intentional activity – as well as Broome's (2013) view that intentions either result from automatic process *demanded by rationality*, or as the conclusion to a practical reasoning process – I take intention to only ever be a possible output when System 2 oversight is present. The intimacy of intention's relationship to System 2 is apparent if we consider System 2's role in long term planning and holistic decision-making, both qualities that are central to the role of intentions. If an intention is to serve its role in diachronic reasoning, it must be possible for it to be incorporated into an agent's general plans, it must be available for comparison with the agent's other intentions, her beliefs, and her desires. Without System 2 oversight these functions would be impossible.

How then to understand how intentions relate to intentional activities? The first step is to recognise that it *is* often the case that intentional activities do follow from intentions. Indeed, this may be the norm. Our System 2 oversight mechanism yields an intention, which then serves as the basis for the guidance of the intentional

---

<sup>47</sup> Bratman emphasises the pro tanto nature of this requirement because he does not wish to rule out the possibility that there may be instances where this requirement may be overturned by some more significant requirement. He does not discuss what such a requirement might be.

activities that follow.<sup>48</sup> This, of course, sounds like Bratman's notion of the motivational potential of an intention determining which activities count as intentional, and it should. I take Bratman to be entirely correct in urging the abandonment of the assumption of tight fit, and adopting a wide-fit view of the relationship between intentions and intentional actions. In effect, whenever an intentional activity has a related intention, the conditions for positive intentional status can be simplified to:

**Intentional status transfer:** an agent's X'ing, undertaken at time t, is intentional if:

- (1) At time t the agent X'd
- (2) X follows, while under System 2 Oversight, from the agent's intention to Y
- (3) The agent has a belief that X'ing is part of the fulfilment of Y

Importantly, X and Y can be the same, but need not be. This point is best illustrated by an example: I have the intention to run a marathon. I then run the marathon guided by this intention and with the belief that it is in fulfilment of my intention to do so. As such, this action can be said to fulfil the motivational potential of my intention to run the marathon. Therefore, I ran the marathon intentionally. On my view, the purpose of an intentional activity in these cases is explained with reference to the intention, which shares positive intentional status with the relevant action. So, in the statement: "I am typing this sentence with the intention of finishing my thesis," the action described (typing the sentence) is an intentional action, while the intention described (finishing my thesis) is the intention that circumscribes the action. Importantly, it is not necessary that my intentional activity to X be related to some intention to X, it must simply be related – through the mechanisms I have described – to *an* intention. So, for example, say I take a drink of water, and that I do so intentionally. My account does not require that I must have had an intention to drink the water, but it is necessary that I must have an intention toward something or other, of which I believe drinking the water is a means to its fulfilment. In this case, quenching my thirst for example.

---

<sup>48</sup> Worth noting again is the fact that this relationship between intention and intentional activity need not be thought of as a causal one.

However, contra Bratman, in at least some cases an intentional activity can take place without any intention having been formed, as was seen when we considered the fringe cases in Chapter 1: Section 3. This usually happens when the activity is very brief, sudden, and uncomplicated – as in the case of certain automatic actions. My account’s explanation for such cases is that the operation of the System 2 oversight mechanism is able to bring about an activity without going through the process of first bringing about an intention. In these instances, there may be insufficient time for an intention to form, and so the agent’s reasons-responsive mechanism results in an action with an intentional content determined by the desired aim and the reasons responded to by the mechanism, but without the mental state of intention (be it conscious or not) or the activation of the related rational requirements.

Without intending to provide a full account of intention, it is worth noting that this is not to say that intentions are only formed by the operation of System 2. Intentions can be the results of System 1 processes, but unlike with intentional actions, not only must System 2 oversight be present – i.e. it must be the case that System 2 had the capacity to intervene or guide the process resulting in the intention – but it must also be the case that this intention be available to the operation of System 2 on demand at the time of acting.<sup>49</sup> Further, for an intention-for-the-future to play its expected role in

---

<sup>49</sup> What this requirement rules out are the kinds of unconscious intentions that Mele (2009: 98-103) discusses when considering an example of table-turning. In this example, a group of individuals are gathered around a table, all of whom believe that the table will move due to spiritual intervention. As it turns out the table does move. As it turns out:

[n]aturally, the people gathered at the table are moving it. But apparently, at least some of them are contributing to its motion without having any idea that this is so. Imagine that one of them, Tab, begins to feel some clockwise motion of the table. His hands move in the direction of the motion, as he notices, and he thinks he is merely allowing them to be dragged along by the table. In fact, however, he is pushing the table in that direction ever so slightly. (*ibid.*: 98-99)

Mele takes this to be a case where Tab intentionally pushes the table, and furthermore that it makes sense to say that Tab has an intention to push the table. I disagree that this is the case. Consider that if there was some consequence of pushing the table that would later inconvenience Tab – that he be accused of being a charlatan for example – he would seem entirely justified and correct, even after being told the details of the case, in declaring, “but I didn’t move the table intentionally! I had no intention of doing so!

This does not mean that intentions must always be conscious, after all I do not lose my intentions when I sleep, but the intentions must be available to my consciousness on demand at the time of acting. Had Tab had an intention to push the table but *forgot* about it until he was in the middle of pushing the table his pushing would then count as intentional if he continued to maintain his pushing, but if he stopped upon this realization then his pushing up till that point would not be intentional. In contrast, if Tab did not forget about this intention of his, but it was simply not “before his mind” at the time, and he could call it to mind on command, then the pushing would be intentional – if the actual pushing did follow appropriately from this intention, of course.

coordinating further intentions and behaviours, it is necessary that it precede the actual further intentions and intentional action in question, and it is also the case that once such an intention is arrived at it immediately evokes the related rational requirements. It is this step that can at times be “skipped” – both due to time constraints, as mentioned before, but also due to increasing degrees of automaticity. In the case of an over-learned action, for example, the intention is not necessarily skipped due to temporal concerns, but because the agent has “ceded” control directly to the System 1 processes resulting in the actions. This is usually unproblematic, as the intention-for-the-future’s coordinating role is often no longer necessary in this case, though it can lead to some difficulties: consider a case where a professional tennis player is teaching a new student. The student delivers a weak shot. In terms of the professional’s overall intentions and beliefs, she would want to strike the ball so as to make it easy for the student to return it again. However, the professional has been rigorously trained and frequently practiced smashing weak deliveries such as the one she has just received, and proceeds to do so. Presuming that in this case System 2 still had oversight – it was not a wholly automatic action – then this smashing of the ball is intentional. Had there first been an intention-for-the-future, System 2 intervention to prevent the smashing of the ball would have been more likely.

An easier way to understand this may be to adopt Searle’s solution to this messy tangle, which is to argue that intention-for-the-future can be divided into two types, prior intention and intention in action. Prior intentions are not always present in cases of intentional action. Intentions in action, on the other hand, are. Searle (1980: 52) distinguishes the two as follows:

The characteristic linguistic form of expression of a prior intention is “I will do *A*” or “I am going to do *A*”. The characteristic form of expression of an intention in action is “I am doing *A*”. We say of a prior intention that the agent acts on his intention, or that he carries out his intention, or that he tries to carry it out. But in general we can’t say such things of intentions in action, because the intention in action just is the Intentional content of the action; the action and the intention are inseparable.

Though I take Searle's explanation to be generally accurate, I disagree that intention in action is a type of intention-for-the-future. Instead I take it to be the aim that the System 2 oversight mechanism was orientated towards. To understand this, consider the following: I am climbing a mountain and suddenly lose my grip and begin to fall, as I do so I grab the climbing rope that I had set up earlier, ending my fall. In this case, there is insufficient time to develop a full-blown intention-for-the-future, instead the aim of grabbing the rope – the intention in action – guided my action in grabbing the rope – without engaging all of the rational apparatus that the formation of a prior intention would have called for. This then accounts for the unity of intention-for-the-future and intentional action.

I take explaining intention-with-which to be simple matter of explaining an intentional activity by referring to the intention-for-the-future to which it is related. When I state: "I am typing this sentence with the intention of finishing this chapter," I am relaying to the listener that I am performing an intentional action (typing the sentence) as part of the fulfilment of an intention-for-the-future (finishing the chapter). This method seems to apply to all uses of intention-with-which that I can presently imagine.

This then gives us the unity of the three uses: both intention and intentional activity are the direct results of an appropriate control mechanism, sometimes together and sometimes not. When there is an intention to A this intention can transfer positive intentional status to a given activity B (where A and B can be the same) if this activity takes place under the intention's guidance (which bottoms out as System 2 Oversight) and with the appropriate belief. In cases of intentional activity where there is no such intention present, we can still think of their being a Searlean intention in action at work, where this intention in action is constitutive of the activity and is identical to its "intentional content" or aim. Finally, intention-with-which is understood to be an explanation of an agent's intentional activity in light of her intention-for-the-future.

Let us now examine how this conception of the unity of the three uses fares in resolving a key difficulty that have plagued previous attempts to reduce the three uses to intention-for-the-future, resolving the problem of mutually exclusive intentions.

### 2.3. Making sense of mutually exclusive intentions

If we say that there is often a direct link between an intentional action and an intention it aims to fulfil, then we must again confront the possibility of mutually exclusive intentions. On the other hand, if we make the link too tenuous, then we run into a different difficulty: explaining the rational relation between intention-for-the-future and intentional action. I believe that the key to plotting a course between these two pitfalls is to give a greater role to uncertainty, by again making use of the notion of *trying*.<sup>50</sup> As I defended in Section 1.3., I take the belief necessary for intentional action to not be a belief that the agent will A, but rather that the agent will at least try to A.<sup>51</sup> The value of this is that it recognises the inherent uncertainty entailed by a future-directed commitment. I will now see how this approach fares in resolving the problem of mutually exclusive intentions.

Tackling Bratman's much-discussed case of the gamer attempting to strike two targets, with the incorporation of the notion of trying we can say that the gamer had the intention-for-the-future to hit one of the targets. This means that the agent has a conduct-controlling pro-attitude toward hitting one of the targets, as well as the belief that she will at least try to fulfil this aim. Does the agent have an intention-for-the-future to hit T1 (or T2)? I argue yes, the agent does. One of the rational requirements applicable to my intention to hit one of the targets is that I should intend those means that would be necessary for trying to successfully fulfil my end. When the gamer decides to set out the two targets simultaneously, she does not know which one she will hit. Even as she is playing, she cannot be certain as to which target she will strike, if any. It could actually increase her chances of fulfilling her overall intention-for-the-future of striking one of the targets if she were to try to strike both simultaneously. In this case, it could be argued that the gamer is following the *most rational* path, as she is maximising the chances of fulfilling her end. At the very least, by aiming to strike both targets she does not reduce the likelihood of fulfilling her

---

<sup>50</sup> For a different approach to mine that also employs the notion of trying, see Thompson (2008: 91–92, 133–146).

<sup>51</sup> It might be worried that incorporating the belief to try as a constitutive part of an intentional action commits me, as Hornsby (1980) has argued, to the view that all basic actions necessarily occur inside the body. Whether this is so does not undermine my account, but I do take Hornsby's conclusion to be premature for reasons presented by Steward (2000) in her compelling criticism of Hornsby's argument. I will merely direct the reader to her work on this point.

overall intention. However, it would be irrational to act as the gamer in the example acts if playing the two games simultaneously makes it impossible to hit either one of the targets (or even if it makes it less likely that the gamer will succeed in hitting one by *trying* to hit the other). And in such a case we would, I think, consider her behaviour to be a case of criticisable irrationality.

Could one not claim, however, that I must still, in the process of meeting the most rational means to the overall intention, hold an irrational combination of subordinate intentions? Perhaps it could be argued that although my overall plan is not irrational, the individual intentions are in fact so. We have seen that Bratman takes this possibility seriously, and attempts to resolve it with the introduction of the notion of “settled objectives” (Bratman, 2009a: 18-19). According to his solution the gamer does not have the intention to hit T1 and the intention to hit T2, but rather has these as settled objectives. These settled objectives do not have the same rational requirements that intentions do. Most relevantly they are not required to meet intention consistency, the very requirement that mutually exclusive intentions violate. Because of this, the fact that these objectives are mutually exclusive is not a concern. Also, any subplans I might form to achieve these objectives would still count as intentional. For example: “I intend to press the fire button to hit target T1.” By distinguishing settled objectives from overall intention in this way, Bratman hopes to avoid the problem of rational contradiction

As I stated before, I take this approach to resolving the problem of mutually exclusive intentions to be unnecessary. Recall that the rational requirement Bratman takes to be at stake here is Intention Consistency, which contends that it is always pro tanto irrational to intend A and intend B, while believing that A and B are not copossible. Notice that it would only be irrational according to this requirement to hold an intention-for-the-future to hit T1 while also holding one to hit T2, if by intention we meant that the agent will succeed, or believes he will succeed. If, on the other hand, we treat seriously the idea that in these cases the agent has the belief that she will *try*, then it is no longer a matter of irrationality. There is no irrationality in *trying* to do two things of which only one can succeed, provided that the agent does not know which one will succeed. In this case, the agent is simply “hedging her bets” so to speak. This does not mean that there are not cases where holding two intentions might

be irrational. If *trying* to do A would make doing B impossible, then I could not rationally intend A and B. In the example of the gamer, if she were to succeed in hitting T1, and in the aftermath still had the intention of hitting T2, then she would be guilty of criticisable irrationality. However, this will not occur if we correctly allow for the role of intention-with-which. My intention-for-the-future to hit T1 (or T2) is subordinate to my intention-for-the-future to hit “one of the targets.” That is, I hold the former intended action as means toward the latter. This is then obviously a case of intention-with-which. If the latter is achieved, then the reason for holding the former is gone. This means that if I intend to hit T1, with the intention of hitting one of the targets, then this intention to hit T1 only makes sense as long as I have not hit one of the targets. The moment I do strike a target, the intention-with-which is either fulfilled (if I struck that target) or unnecessary (if I struck the other target). In either case, I will not continue to hold the intention after the point of fulfilment, as the reason for holding the intention is no longer there. This solution retains a direct link between the relevant intention-for-the-future and the relevant intentional action and avoids having to introduce the notion of settled objectives.

Another possible solution – one that does away with the idea that hitting T1 and hitting T2 are intentions-for-the-future – goes as follows: I have a single intention-for-the-future, to hit one or other of the targets. I have two intentions-with-which, each related to either T1 or T2 respectively. The intention-with-which related to T1 would be, “I am trying to hit T1 with the intention of hitting one of the targets.” Obviously, the action of trying to hit T1 would then be an intentional action, and my intention-for-the-future is still “hitting one or other of the targets.” This approach resolves the issue by avoiding us having to say that I have the intention-for-the-future of hitting T1 (or T2) specifically, and so avoiding the risk of irrationality. It also avoids having to introduce the notion of settled objectives. I think either of these explanations is acceptable, and that adjudicating between them would require asking the gamer what her intentions(-for-the-future) actually are, that is to say, I think both kinds of cases are possible.

## 2.4. Concluding Remarks

It has been my aim in this chapter to present a cogent and compelling account of intentional activity and intentional outcomes. I contended that this should be a control-backed account, where the presence of intention-unique control is taken to be the characteristic feature of intentional behaviour. I also introduced the idea that we should start our attempt to provide the aforementioned account by investigating the actual mechanism via which an agent exerts their rational agency in behaviour. In order to achieve this, I first presented the requirements that must necessarily be met by such a mechanism: that it be reason-responsive, that it explain the possibility of acting against the balance of reasons, and that it has the ability to result in both intentional activities as well as intentions. This done, I then introduced Dual Process Theory, and argued that the sought for mechanism is best understood as System 2 oversight. I then showed how this understanding of the intention-unique control mechanism as System 2 oversight meets the three requirements I introduced. This lead into my presentation of my own initial formulation of the control condition on intentional action: *System 2 Oversight*.

In the three sections succeeding this, I showed how my initial formulation was inadequate, first by explaining the necessary role played by a belief to try in meeting the epistemic condition on intentional behaviour. And secondly, by arguing that rather than focussing only on intentional *action*, a proper account of intentional behaviour should expand to consider intentional omissions, effects, and consequences. I then argue that the intentional actions and omissions can be considered together under the label of intentional activities, given that both are abstractions (or activity-slices) individuated from the stream of activity. Intentional effects and consequences I labelled together as intentional outcomes, though unlike with activities these two come apart in a significant way. Whereas intentional effects (those results of intentional activities that are their logical consequents) can be properly accounted for by a reformulation of S2O – System 2 Oversight\*\* – intentional consequence derive their positive intentional status via a transfer principle: *Transfer of intentional status*. This difference, I contend, follows from the different degrees of control that an agent can exert of intentional effects and intentional consequences respectively.

With my account sketched, I then sought to bolster the case in favour if it by seeing if my account could meet the Unity Criterion, by explaining the relationship between the three uses of the concept intention famously identified by Anscombe. Having established my account's compatibility with the unity of the three uses, I then proceeded to argue that my account has the necessary tools to resolve the problem of mutually exclusive intentions. This is achieved by recognising that *trying* is constitutive of intentional activity, though this does not mean that “intending to X” is equivalent to “trying to X”, but rather that intentionally X’ing necessarily means that the agent must have the belief that she will at least try to X.

Equipped with this account of intentional activity and intentional outcomes – and having adduced support for this account by demonstrating its ability to account for the fringe cases discussed in Chapter 1, and to resolve or avoid a problem that has beset the various accounts of intentional action that was discussed there – the next step in untangling the relationship between intentional activity and moral responsibility will be to turn toward the latter notion. This I do in the next chapter.

## CHAPTER 3: MORAL RESPONSIBILITY

### Introduction

Any discussion of the topic of moral responsibility finds itself embedded within a truly vast, complex, and ever-growing literature. Given the enormous scope of the work that has been devoted to just about every aspect of this topic, my discussion of it in this chapter will, by necessity, be a less than complete account. I will be focusing only on those dimensions of the discussion that I deem most relevant for my project: charting the relationship between moral responsibility and the intentional status of some *activity or outcome*, and only in the case of the *individual agent* – so excluding responsibility for *attitudes*, or considerations of *collective responsibility*. This will undoubtedly mean that several aspects of the wider discussion will be passed over in a manner that some will find unsatisfactorily brusque – and rightly so. However, I take this to be a regrettably inevitable result of wading into such densely populated and hotly contested waters. We must strive to untangle one knot at a time, and the knot I am seeking to unravel here can be understood as the question, “what is the relationship between the intentional status of an agent’s activities and outcomes, and the agent’s openness to moral responsibility for these activities and outcomes?”

Answering this question, as I have already mentioned, involves first getting a clear account of both the central notions involved: intentional activities and outcomes, and moral responsibility. I take the foregoing two chapters to have provided an account of the former, and take this chapter to provide an account of the latter. Thus, this chapter seeks to provide a plausible, though not exhaustive, account of moral responsibility for the activities and outcomes of individual agents. To meet this goal, I begin by setting the stage. This involves situating my account relative to two ongoing debates in the literature: the debate between merit-based and consequentialist understandings of moral responsibility, and the complex debate surrounding the relationship between moral responsibility and determinism (and indeterminism, as we shall see). The two debates are not wholly unrelated, as the consequentialist understanding counts as one of its virtues that the truth of determinism would not undermine their position, whereas this is not as obviously the case for merit-based understandings. However, I will discuss each separately to most effectively clarify where my account stands with

regards to the various positions available in these debates, namely: that I am endorsing a merit-based understanding of moral responsibility, and that I am developing my account *as though* compatibilism where true, while maintaining agnosticism about the actual truth of compatibilism – and leaving it open to a non-compatibilist to treat my account as a modular component that can fit into a wider account. By this I mean that it can either be treated as an account of conditions that are necessary but not yet sufficient for responsibility (a Libertarian for example may argue that a further freedom condition of some sort must be added), or maintaining that the conditions for responsibility in my account may be accurate but in fact are never met (as a hard determinist or skeptic of any stripe may argue).

My account's position in the discussion of moral responsibility's wider landscape having been established, I will move on to argue that there is an important (and further: necessary) relationship between the intentional status of a behaviour and moral responsibility. This being the case, and since I presented a *control* account of intentional behaviour in Chapter 2, it is likely unsurprising that I will argue that it is through the *nexus of control* that these two notions interrelate. This puts the role that control plays in moral responsibility on center stage. This is a comfortable position for the role of control to occupy, as the debate surrounding it is very much a central part of the contemporary discussion surrounding moral responsibility. This debate can be understood as one between *volitionists*, who hold that control is necessary for moral responsibility, and *attributivists* (most relevantly for my purposes Quality of Will theorists), who argue that an agent is responsible for something insofar as it appropriately reflects some relevant quality of the agent herself (where what is meant by “relevant quality” is, as we will see, offered various interpretations). Crucially, attributivist accounts contend that the presence of agential control over some behaviour is not a necessary condition for that behaviour to be reflective of that quality of the agent that determines openness to moral responsibility.

I will examine each of these two approaches in turn, before presenting what I take to be a problem that confronts both: the existence of cases that elicit ambivalent moral responses. This is not a new problem, but one that was raised by Watson (2004) and refined by Shoemaker (2015a). In brief, the problem highlights the existence of cases where it seems that an agent's behaviour justifies certain moral responses but not

others. Importantly this is not a difference in *degree*, but a difference in *type*. It is not that we doubt whether an agent is morally responsible in these cases, but rather that we are *ambivalent*, often strikingly so, about the type of moral responses that we are legitimated to adopt. This is a difficulty for volitionist and attributivist accounts as they offer invariantist accounts of moral responsibility, where moral responsibility may come by degree, but not by type. Following Watson and Shoemaker again, I will be arguing that the existence of these cases should motivate us to recognize that moral responsibility is a *plural*, rather than unitary, notion. In other words: that there can be different senses or variants of moral responsibility, and that each of these may have differing conditions they must meet to obtain, and in turn that they justify different moral responses. In my arguments here, I will only posit two variants of moral responsibility: *attributability* and *accountability*. Amongst other differences that I will discuss, my crucial contention is that the presence of control is a necessary condition on the latter, but not the former. Relatedly, the presence of moral attributability legitimizes a different set of moral responses than does the presence of accountability.

Having sketched these two variants, I will then argue that, to some extent, volitionists and attributivists have been talking past each other. They have been talking about different variants of moral responsibility. I take attributivist accounts to be focused on the notion of moral attributability, whereas the volitionists seek to limit moral responsibility to accountability. Each of these positions has something right and something wrong about them: the attributivists are correct to identify that there is a variant of moral responsibility that extends beyond the limits of control, but are wrong to think that there is not another substantive variant of moral responsibility applicable in those cases where appropriate control is present. Volitionists are correct to identify that certain of our most important moral responses are only legitimated in cases where control is present, but wrong to think that the kind of responsibility present in cases that extend beyond controlled behaviour is not also a variant of moral responsibility. By adopting a variantist view of moral responsibility, this tension can hopefully be softened, even if not resolved.

Turning the dialectic back toward my overarching argument, I will contend that it is moral accountability for behaviours – due to its unique relationship to control – that stands in a necessary relationship to the intentional status of a given behaviour. To

prove this, I present what I take to be the control condition on moral accountability for activities and effects: *normal System 2 Oversight*. As the name suggests, this kind of control is intimately linked to, but not identical with, the control necessary for positive intentional status. Normal S2O is more restricted than S2O. In particular, normal S2O demands not only that a behaviour be under System 2 oversight, but that the dual system mechanism in question be *functioning normally* – which includes the requirement that the mechanism be sufficiently responsive to moral reasons. The use of “normally” here is not colloquial, but technical, relating to the dual system mechanism’s operative characteristics (Haldane, 2011), and the idea that an abnormal state of affairs is one that demands a *special explanation* (Smith, 2010: 15). As we will see, this restricts the scope of behaviours for which an agent can be morally accountable as compared to those that are intentional. In addition, agents can be morally accountable for intentional outcomes, even though such outcomes are not *directly* under normal System 2 oversight. To capture this, I introduce *normal accountability transfer*, whereby an agent can be accountable for a given consequence provided, amongst other requirements, that the consequence was *sufficiently under the guidance of normal S2O*. Having outlined the control condition on accountability, I will argue that an activity only meets normal S2O when S2O is met, and that because of this an agent is only ever morally accountable on the basis of her activities if they are intentional. And furthermore, that this relationship extends to outcomes as well. This is what I term, the *nexus of control*.

I then turn to a discussion of some of the consequences of my account for cases involving the Doctrine of Double Effect, and finally return to the inciting example of Truman and Anscombe.

## 1. Setting the stage: positioning my account

### 1.1. Merit versus consequentialist accounts

In this chapter I will be embracing a *merit-based* understanding of moral responsibility, as opposed to a *consequentialist* one. These two positions have, until fairly recently, dominated how the traditional discussion of moral responsibility has been conducted. According to the desert-based view, to hold an agent morally

responsible, in the sense of either moral blame or praise, would be an appropriate reaction toward the candidate if and only if she “merits,” or “deserves,” such a reaction (King, 2014: 1). The consequentialist view holds that such praise or blame would be appropriate if and only if a reaction of this sort would likely lead to a desired change in the agent and/or her behaviour (Schlosser, 2013: 226), or in the behaviour of others. In this case, what it is to be morally responsible depends upon considerations of what moral blame or praise will achieve in a given case. Consequentialist accounts derive not inconsiderable appeal from the fact that they seem more robust than merit-based accounts when confronting the dual threats of determinism and moral luck. Nothing in the consequentialist account *necessarily* requires that an agent that is held morally responsible for some action need have been able to act differently than she or he did, or have been in control of why and how she or he acted. Such considerations may play a role if they influence whether or not holding the agent responsible would bring about the desired change, but not otherwise. Merit-based accounts, on the other hand, must confront these potential threats more directly, either by denying them and arguing that they are in some way untrue or mistaken (in the case of Libertarians) or arguing that they do not represent a threat to moral responsibility (in the case of compatibilists). I will be discussing the incompatibilism/compatibilism debate, and where my account is positioned in relation to it, in more detail in the next section, for now it is sufficient to know that consequentialism about moral responsibility allows a proponent to sidestep this debate, whereas this course is not open to a supporter of the merit-based approach.

Consequentialism is not very widely endorsed, certainly not in the current discussion. For most it simply seems misguided to think that an agent’s openness to moral responsibility should depend upon the future-directed value of holding them responsible. And adopting this view is radically at odds with our actual practices of moral responsibility. Though such considerations might deserve attention when asking the question of whether or not to respond in certain ways to those who are responsible, it is peculiar to think that this determines the actual presence of responsibility. This is in no way a knockdown argument against consequentialist accounts, and it is not my aim within this piece to provide one.<sup>52</sup> I will, however, be

---

<sup>52</sup> For some influential arguments against consequentialism see Strawson, 1962; Wallace, 1994; Scanlon, 1986.

following the recent trend in the literature in taking a merit-based view of moral responsibility.

### *1.2. Incompatibilism and compatibilism*

Related to, but distinct from, the contrast between merit-based and consequentialist accounts, is the venerable debate between incompatibilists and compatibilists about moral responsibility. These positions hardly need introduction, and the influence of this debate has been so pervasive, that any serious attempt to unpack moral responsibility – at least any merit-based one – must clarify its position on the matter. Traditionally, the central concern of this debate was the possibility or impossibility of free will, and what consequence this has for the possibility of moral responsibility. As expressed by Mason (2005: 344):

[When considering moral responsibility i]t is impossible to ignore the issues of free will and determinism entirely. Moral responsibility is a problem because our best theories about our physical world tell us that our actions are caused by mechanistic processes originating outside of us. It seems that if our actions are ultimately caused in the same way that avalanches are caused, then ultimately we are no more responsible for what we do than a rock is for knocking into another rock. The problem has led some philosophers to argue that determinism is false, and that either agents are special causes (and thus morally responsible for what they do) or that indeterministic events can account for moral responsibility.<sup>53</sup>

Of course, there are also those philosophers who conclude, based on the assumption of the truth of determinism, that moral responsibility (at least in the way we commonly conceive it) is in fact impossible, or in need of radical revisionism.<sup>54</sup> It is also worth noting that even if indeterminism is granted, there remains a worry that this too may be incompatible with free will – and so by extension moral

---

<sup>53</sup> These approaches are often labelled as *source* and *leeway* incompatibilism respectively. Both these types of accounts put forward a condition on moral responsibility that is intended to capture the unique freedom that they take as necessary for such responsibility: In the case of source incompatibilism this condition is that the agent must be the ultimate source of her behaviour and in leeway incompatibilism the agent must have access to metaphysically robust alternate possibilities (Kane, 2002; 2007 and Haji, 2002)

<sup>54</sup> Examples include Galen Strawson (1994) and Derk Pereboom (2007; 2013).

responsibility, as it then seems that outcomes are random, and that such randomness rules out responsibility. Not that this worry cannot be potentially responded to, perhaps by limiting and specifying the degree and type of randomness at work – as Levy (2011: 1-2) and Kane (1996: 2007) argue may be possible – but it remains a spectre haunting attempts to defend the possibility of free will and moral responsibility by introducing indeterminism into the picture. However, perhaps the most common response amongst philosophers has been to pursue compatibilist strategies, which hold that the truth of determinism does not threaten the possibility of free will, or that even if it does, it does not threaten the possibility of moral responsibility.<sup>55</sup> I will drop the distinction between these two types of compatibilists from here on, as this distinction is not relevant to either understanding the dialectical position of my account of moral responsibility, or to my overarching argument.

Compatibilists of either stripe have tended toward talk of *control*, or *self-disclosure*, rather than free will itself. Taking the first approach leads to the view that what is necessary for moral responsibility is the presence of a certain kind of agential control. In contrast, the second approach maintains that what is necessary for an agent to be open to moral responsibility for some behaviour is that that behaviour must disclose something morally relevant about the agent – variously argued to be the agent's character, “real” or “deep” self, or the agent's quality of will or care. It is already worth noting here that approaches that hold control to be a necessary condition are also concerned with self-disclosure in the sense that they argue that only behaviours that are under the appropriate kind of control can reveal facts about the agent's morally relevant features (at least insofar as blame and praise are concerned). As will be discussed in more detail in Section 4, whether control or self-disclosure is taken to be the central requirements for moral responsibility is a topic of great importance in determining the relationship between moral responsibility and the intentional status of behaviours. For our purposes here, it is enough to provisionally identify these two types of compatibilism.

More recently, the argument has been advanced that it is not determinism or indeterminism that renders free will and moral responsibility impossible, but rather

---

<sup>55</sup> Fischer and Ravizza's (1998; 2007) account is an example of the latter view, which is why they identify their position as one of “semi-compatibilism.”

the presence (and indeed omnipresence) of *luck*. This view, whose most eloquent proponent is probably Neil Levy (2011), holds that the presence of such luck undermines the possibility for the kind of control necessary for moral responsibility, as well as problematizing self-disclosure accounts by questioning how the revealed morally relevant features of an agent can be a basis for moral responsibility if (i) which features an agent possesses, (ii) if they are able to reveal them, and (iii) whether any given behaviour is a reliable guide to these features, may all be a matter of luck. Additionally, he contends that the epistemic condition on moral responsibility is far more stringent than is commonly thought, and that in few, if any, cases does an agent meet them. For Levy's argument to get off the ground, it is first necessary for him to convince us that the conditions for moral responsibility that he puts on the table are accurate, as whether we should agree that luck undermines these conditions will depend on what they are. I will directly consider these conditions in Section 3.1., as in that section I discuss volitionist accounts of moral responsibility, of which Levy provides an example (though in the end he concludes that these volitionist conditions are in fact never met).

Having briefly charted over the landscape of this debate, I can position my own account's place in it: I will be developing a partial and modular account of moral responsibility that I develop *as though* it was a compatibilist one. This account will be *partial*, as I do not, and do not intend to, provide a defence of compatibilism as such. Nothing that I present here will convince an incompatibilist (be they a skeptic or a Libertarian) of the truth of compatibilism. Rather I will provide what I take to be a plausible and compelling understanding of moral responsibility if we assume the truth of compatibilism. What is true is that I take my account to be amenable with these incompatibilist positions in the sense that it is *modular* enough to accommodate them: nothing in my account rules out the possibility that the conditions on moral responsibility I argue for never obtain (as a skeptic may argue), or that on top of the conditions I put forward there be some further demand for sourcehood or leeway (as Libertarians may argue for) for moral responsibility to be present. I will be agnostic on this matter. Hopefully this modularity and agnosticism can allow something of interest to be gained by incompatibilist readers.

## 2. Moral responsibility and intentional action

In this section I give a preliminary look at the relationship between moral responsibility and intentional action. This will serve to demonstrate the complexities involved, as well as indicate to us the appropriate potential course to resolve them. The first step is to consider what it is I mean by *moral responsibility, praise, and blame*.

When attempting to understand what it means for an agent to be morally responsible, one popular general view is that “[w]hen we talk about moral responsibility we are talking about a way of being related to things so as to make individuals [morally] blameworthy or praiseworthy for them” (King, 2014: 2). According to this general view, what we mean when we say that an agent is morally responsible is that the agent is a legitimate candidate for ascriptions of (moral) praise or blame. By the same lights, to hold an agent morally blameworthy or praiseworthy is to hold the agent to be morally responsible.<sup>56</sup> More recently, Strawson, in his landmark paper, “Freedom and Resentment”, introduced the idea that what it is for us to hold an agent morally responsible, in the sense of blameworthy or praiseworthy, is to adopt certain reactive attitudes toward them, attitudes such as resentment, indignation, hatred, amongst others. However, I want to take a moment to consider an important distinction that arises from the consideration of the intimate relationship between moral responsibility and moral praiseworthiness and blameworthiness. Although it is the case that it seems uncontroversial to claim that to hold an agent to be morally blameworthy or praiseworthy is to hold her to be morally responsible, this does not yet allow us to say that *what it is to be* morally responsible, is simply to be morally blameworthy or praiseworthy. The reason for this, as King (2014: 2) observes, is that it can be thought that an agent could be morally responsible for an action but “fail to be either blameworthy or praiseworthy for it. Presumably this occurs when one is responsible for a morally neutral act.” In cases such as this it seems that an agent can be morally responsible even for actions for which she is neither praiseworthy or blameworthy.

Fischer and Ravizza (1998: 8) endorse this view when they declare that:

---

<sup>56</sup> This approach has a long pedigree: Aristotle adopted it in his discussion of moral responsibility in *Nicomachean Ethics* III. 1-5.

[M]oral responsibility need not imply the actual application of a reactive attitude; it only requires that the agent be an apt candidate for such an application...[i]n deed, our Strawsonian view of moral responsibility allows moral responsibility for “morally neutral” behaviour. For instance, one can be morally responsible for simply raising one’s hand (where this is not a signal or in any way morally significant). Thus our Strawsonian view of moral responsibility is a relatively broad and inclusive view.

I am, according to this view, morally responsible for painting my garage door, but barring exceptional circumstances I am unlikely to be morally blameworthy or praiseworthy for it. It is this understanding of moral responsibility and its relationship to praiseworthiness and blameworthiness that I will be adopting.

Now, it is very often the case that a denial of positive intentional status exculpates an agent from moral responsibility for some action. Indeed, such protestations are abundant in everyday life. And what is more noteworthy than the fact that these denials are advanced, is that they often do serve to render the relevant attributions of moral blame or praise unjustified. If we imagine the following case: a man in a crowd treads on my foot, and I angrily blame him for this, but if he responds by removing his foot while truthfully stating, “that wasn’t intentional,” then it seems eminently plausible that we will find it to have been unjustified of me to hold the agent morally blameworthy for his treading on my foot.<sup>57</sup> The same effect operates in the opposite direction: if an agent denies moral blameworthiness for some act, it usually counts against this denial to say that the agent in fact performed the action intentionally. These everyday practices might motivate someone to advance a strong answer to the question, “is there a necessary relationship between an agent’s openness to moral responsibility for some action and the intentional status of said action?” (call this question Q1), namely:

**A1:** an agent is only open to moral responsibility for her or his intentional actions

---

<sup>57</sup>Of course, the matter changes if, once made aware of my foot under his, the agent refuses to move his offending appendage. However, I take this to only reinforce my point: by refusing to move his foot after having been made aware of the situation, the agent can no longer claim to be treading on my foot unintentionally.

However, this is quite clearly too strong an answer. It attempts to capture a far more complex reality under an overly simple rule.<sup>58</sup> That said, certain adherents to volitionism – such as voluntarists – do endorse something like A1. As McKenna (2008: 31) points out, “voluntarism about moral responsibility is a thesis about control and its scope, which is limited to intentional actions.” However, reconciling A1 with our actual practices of moral responsibility introduces complexities that make such a straightforward answer untenable. There are at least two good reasons to reject A1: (i) there are at least some cases where an agent can be morally responsible for actions which are not intentional and (ii) agents can be morally responsible for more than merely actions.

To illustrate this, consider the following case:

Mary is a tourist, and she is busy climbing across a snow-covered slope at a resort when her leg becomes stuck in a snow drift. Mary has with her a flare gun, which she has been instructed to fire if she becomes stuck. However, she has also been warned that she should avoid firing the flare when she is stuck on parts of the eastern slope as this could cause an avalanche. Mary is in fact on the eastern slope, but the warning slips her mind as she is in a panic to save her life – an end she values highly – and she fires her flare gun to signal for a rescue ignorant of this fact. This has two consequences, the first is that Mary is indeed rescued, the second is that an avalanche is caused which strikes a town in the valley below killing a dozen people.

In this case, it seems that we would hold Mary to certainly be morally responsible in some sense for the deaths that resulted from her actions, yet it cannot be said that Mary *intentionally* killed the townspeople. So, in this case, the defence of “it was not intentional!” does not seem to be compelling. There are myriad similar examples that could be put forward, many of which come from everyday life: if I agree to housesit a friend’s dog, but have an apathetic attitude towards my duties and absentmindedly feed the animal chocolate, its harmful effects on dogs slipping my mind, and it dies,

---

<sup>58</sup> The observation that there are cases where an agent might be morally responsible for an action even if that action is unintentional can be found at least as far back as Aristotle’s discussion of the relationship between the voluntariness of an agent’s action and the openness of said agent to responsibility on account of that action in the *Nicomachean Ethics*. Though he discusses *voluntary*, rather than *intentional* actions, the insight he makes applies as readily to the latter as the former.

then it cannot be said that I intentionally killed the dog, however it would be strange to say that I was not morally responsible for it.<sup>59</sup> That said, it is worth noting even here that these cases are likely to elicit deep ambivalence as to what exactly the moral responses applicable in these situations should be understood to be. After all, the way in which Mary is morally responsible for the deaths – both in terms of what it tells us about her agency and what *responses it seems to morally legitimate* – is different not merely in degree from a case where Mary intentionally caused these deaths. We will return to this point in Section 4.1.

As regards (ii), one might simply point out that it is very common for agents to be held responsible for reprehensible attitudes, even if these attitudes never manifest into action. It is surely legitimate to hold a racist morally responsible for her racist attitudes, even if she never performs any harmful actions because these attitudes. However, I will be ignoring responsibility for attitudes in this dissertation, and focus only on behaviours, given that my interest is the relationship between moral responsibility and intentional behaviour. More relevantly then: agents can be responsible for inaction just as they are for actions, as is usually the case with omissions and negligence.

I take negligence to be criticisable omission, though this criticisability need not be moral in nature. Following the arguments of John Haldane (2011: 617), I take an omission (by a given agent) to X, and so leading to a result Y, to be a case of that agent not doing X when (a) it was in her power to X, (b) to X was one of her operative characteristics, and (c) under normal conditions had she X'ed then Y would have been avoided. By operative characteristic here is meant some role or function (partly) constitutive of the agent. Such characteristics give rise to normative expectations regarding the agent's behaviour: if the agent is a gardener, for example, then to water the plants would be one of his operative characteristics, and there would be a normative expectation on the gardener to do so (*ibid.*: 617-618). For an action to be negligent then it must be the case that Y is a criticisable outcome (again, this need certainly not be moral criticism), and on the basis of this the omission to X which

---

<sup>59</sup> There are also many similar, if usually more contrived, examples in the experimental philosophy literature that indicates that the folk are very prepared to ascribe moral responsibility to agents on the basis of actions (or the consequences of actions) that are not intentional, see Malle and Knobe, 1997; Doris *et al.*, 2007; Woolfolk *et al.*, 2006; Nado, 2008.

results in Y coming about is also criticisable. For example, the gardener could have the responsibility of ensuring that the grass remain cut, and so would be negligent if she omitted to mow the lawn, and would be criticisable in the domain of garden maintenance. In the domain of morality, negligence means an omission to act when the balance of salient moral reasons call for action. So, if an agent walks past one of the many drowning children that populate the works of moral philosophers, and if we assume (for the sake of argument) that there are compelling moral reasons to save the child given the low cost to the agent, then if the agent were to fail to act per these reasons and rescue the child, we would hold this agent to be morally blameworthy on the basis of her omission. I think it would be accurate to say that the agent demonstrated moral negligence, just as the gardener in the foregoing example demonstrated gardening negligence. What fixes criticisability seems to be related to the operative characteristics at play. In the case of the gardener, it is the agent's operative characteristics in the role as a gardener that make the failure criticisable. What makes the moral case different (and perhaps unique) is that the role of moral agent is such a near universal one for human agents.<sup>60</sup> One of the operative characteristics of a moral agent is that such an agent should strive to identify salient moral reasons and be guided by them in behaviour, and an agent is criticisable when they do not. The rub of this is that it seems that we have good reason to think that agents can, and often are, morally responsible for omissions as well as for actions.

Another aspect of (ii) is that we at least sometimes hold agents responsible for the side-effects of their actions. This was, of course, a central part of what Anscombe was asserting in her dissent to the Oxford committee on the question of awarding Truman an honorary degree. Though the question of whether an agent is *as* morally responsible for the intended consequences of her actions as for the unintended side-effects can be set to one side for discussion later, what is uncontroversial is that at least sometimes an agent can be morally responsible for an unintended side-effect. Consider a case where a sniper lines up to shoot at his target, but notices that a bystander located directly behind her target will undoubtedly die if she pulls the

---

<sup>60</sup> Depending on one's view of human agency, it may well be taken as universal. For my part I would say that an agent such as an ideal psychopath – who is incapable of recognizing the special normativity of moral reasons – still qualifies as a possessing human agency, but is not a *moral* agent. To identify and respond to moral reasons is not an operative characteristic of a *human agent*. But this matter is neither here nor there for the purposes of my arguments here, though we will return to it in Section 5.1.

trigger. Our sniper has no interest or desire to kill the bystander, and intends only to kill the target. If this sniper were to take the shot, hence killing both her target and the bystander, it seems reasonable to assume that she would be found morally blameworthy for having killed the bystander. Yet this consequence of taking the shot was an unintended side-effect – not even an intended consequence, let alone an intentional action.

Given these complexities, it should be clear that (i) and (ii) give us good reason to reject the thought that the only actions on the basis of which an agent can be morally responsible are intentional actions, and that only actions can be the basis for moral responsibility. Thus, we should reject A1. What then should be our answer to Q1? Before presenting a straight answer to Q1, let us first have another look at the question again:

**Q1:** is there a necessary relationship between an agent's *openness to moral responsibility (of some kind)* for some action and *the intentional status of said action*?

My contention is that to present an appropriate answer to this question it is first necessary that we precisify the two crucial components at work in the question. In other words: what *kind of moral responsibility* is it that stands in a necessary relationship to the intentional status of that based on which the agent is open to responsibility? And the *intentional status of what*? In the preceding two chapters I have already developed my answer to the second question: the intentional status of activities and outcomes. In the succeeding parts of this chapter I argue that the answer to the first is *moral accountability*. Furthermore, I will argue that what grounds this necessary relationship is the *shared role played by control* in both notions – what I will call the *nexus of control*. Having already presented my account of the role of control in intentional activity and outcomes, I now turn to discussing the role of control in moral responsibility. It is through this discussion that we both come to understand why it is moral accountability that stands in a unique relationship to intentional status, as well as the exact nature of the shared nexus of control.

### 3. The contested role of control: volitionist and attributivist accounts

In Section 1.2. a distinction was introduced between two compatibilist strategies: one that takes the presence of some appropriate kind of control to be necessary for moral responsibility, and another that focusses on the revelatory nature of an agent's behaviour. These two strategies have somewhat crystalized in the contemporary discussion into two highly influential schools of thought: that of volitionism and attributivism. Though there are other differences between these approaches, the most important one is how they differ in terms of the role given to the control principle, which holds that for an agent to be morally responsible for something, it must be under the agent's control in the right way (Watson, 2004: 269). Volitionism holds that there is a necessary control condition on moral responsibility. This control condition is usually taken to be composed of two subconditions: the epistemic (or knowledge) condition and the volitional (or control) condition.<sup>61</sup> The former of these is roughly a demand that the agent must know what she is doing to be responsible for it, and the latter is usually understood as requiring that the agent must have the capacity to adjust or guide her behaviour in the light of reasons. Both of these subconditions should sound somewhat familiar from our discussion of intention-unique control in Chapter 2.

By contrast, attributivist accounts argue that there is no *necessary* control condition on moral responsibility. On such accounts, an agent is morally responsible for some activity if it reveals facts about some morally relevant feature of the agent. This feature is variously understood as the agent's “real self” or “deep self” (Watson, 2004), or as an insufficiency or surfeit of *good will* (Rosen, 2014), or how much an agent *cares about what is morally important* (Björnsson, 2017a). Crucially, this can be true *regardless* of whether the agent in question's activity was under his or her control in the volitionist's sense (Radoilska, 2016). However, as we will see, this does not strip control of its role entirely, though it is much reduced – or so the attributivist claims.

---

<sup>61</sup> For the sake of clarity, I will be referring to the overall control condition as the “control condition,” and the subcondition as the “volitional condition.”

In the rest of this section I will expand on each of these approaches to understanding moral responsibility, highlighting some of their strengths, shortcomings, and shared elements.

### *3.1. Volitionism*

To understand the kind of control that I argue volitionists *should* be concerned with, it is helpful to re-introduce Fischer's two types of control: regulative control and guidance control. As a brief reminder, regulative control requires that the agent have access to actual alternative possibilities, whereas guidance control requires that the behaviour be guided by a reason-responsive mechanism. The relationship between the two types of control is well expressed by Sartorio when considering the question of what grounds the kind of control or freedom necessary for moral responsibility (2016: 108):

[T]wo models have emerged as competing answers to this question: the alternative-possibilities model, which is the classical model of freedom, and, more recently, the actual-sequence model (I shall refer to them as “the AP model” and “the AS model,” respectively). According to the AP model, freedom is grounded, at least partly, in having access to alternative possibilities of action. In other words, acting freely consists, at least partly, in being able to do otherwise (being able to do something other than what one actually did). By contrast, according to the AS model, freedom is exclusively grounded in facts about the actual sequence of events issuing in one’s behavior. On this view, acting freely is just a matter of one’s behavior having the right kinds of actual causes, and thus is not at all a matter of being able to do otherwise or having access to alternative possibilities of action.

Regulative control falls in line with the AP model, whereas guidance control is an example of the AS model. If accepted, the truth of determinism seems to clearly rule out the former of these. If nobody has the power to change the course of the future, then nobody has access to alternative possibilities, and so can never claim to have regulative control as an AP account would require. A compatibilist would be wise then to adopt the AS model, rather than the AP one.

A further challenge to the idea that regulative control is necessary for moral responsibility can be found in the famous Frankfurt-examples.<sup>62</sup> Over the years there have been any number of variations on the classic Frankfurt-example, and the one that Fischer (2007: 58) uses to illustrate why regulative control is seemingly not a necessary condition for moral responsibility is the following:

Jones has left his political decision until the last moment...Jones goes into the voting booth, deliberates in the ‘normal’ way, and chooses to vote for the Democrat. On the basis of this choice, Jones votes for the Democrat. Unbeknownst to Jones, he has a chip in his brain that allows a very nice and highly progressive neurosurgeon (Black) to monitor his brain. The neurosurgeon wants Jones to vote for this Democrat, and if she sees that Jones is about to choose to vote for the Republican, she swings into action with her nifty electronic probe and stimulates Jones’ brain in such a way as to ensure that he chooses to vote for the Democrat.

He continues to point out that “[g]iven this set-up, it seems that Jones exhibits guidance control of his vote, but he lacks regulative control over his choice and his vote.” Why would this be? Why does Jones seem to lack regulative control in this case? The reason is that it seems as though he does not have access to alternative possibilities. No matter what Jones may do, he would vote Democrat. He cannot, through any freedom or control internal to himself, shift onto some alternative branch of history. Indeed, it would seem as if, at least in this case, history has no branches at all. Yet, it seems intuitively obvious (or so compatibilists will argue) that we can hold Jones responsible for voting for the Democrat in the actual story (the one where the neurosurgeon never had to step in). The compatibilist should contend that what this shows is that we can hold agents moral responsible for their actions, even in cases where they did not have access to alternative possibilities.<sup>63</sup>

---

<sup>62</sup> Harry Frankfurt first presented examples of this type in his 1969 work, *Alternate Possibilities and Moral Responsibility*. They were intended to disprove the principle that moral responsibility requires that an agent have access to actual alternative possibilities, by positing a situation where the reader intuitively deems the agent to be morally responsible, but the agent *could not* have done anything other than what she did.

<sup>63</sup> Needless to say, this conclusion is not usually meet with agreement from Libertarians or skeptics, see Robert Kane (1996; 2007) for the former and Derk Pereboom (2007; 2009) for the latter.

Recalling from Chapter 2: Section 1.1. that reason-responsiveness as discussed above consists of two elements: reason-receptivity and reason-reactivity, it is worth considering a criticism advanced by Pereboom (2009) of the reactivity facet of reasons-responsiveness, at least as this facet is presented by Fischer. Pereboom (*ibid.*: 29) uses the term “weak reasons-reactivity” to describe the position that there are at least some near possible worlds where the agent’s mechanism is reactive to some relevant reasons. He then proceeds to argue that weak reasons-reactivity is not a necessary condition for moral responsibility. This has primarily taken the form of examples which seem to show that it is possible to imagine an agent who is morally responsible for her actions, yet there is no near possible world in which she could have adjusted her behaviour in the light of relevant reasons to anything other than what she actually did – ergo, weak reasons-reactivity is not necessary for moral responsibility. One such example progresses as follows:

So imagine that someone comes to your door and wants to know whether you are lodging a particular person – and in fact you are. But you are so committed to telling the truth that you would do so under any circumstances, even if you knew that the person at the door was planning to murder your guest, or even if you knew that he would destroy the whole world if you told the truth. You might be morally responsible for your truth-telling despite not being weakly reasons-reactive. Such cases indicate that weak reasons-reactivity is not necessary for moral responsibility. (Pereboom, 2009: 30)

Pereboom does recognise that one route to defusing this example would be to say that the agent in question, A, is in fact not morally responsible for the behaviour, as clearly the kind of non-reactivity evinced from A’s behaviour is more in line with that of an extreme kleptomaniac, or a sufferer of agoraphobia whose psychological limitations are so severe that she cannot leave her house even to save lives. His response is that we should imagine A’s commitment to telling the truth to be “of the sort Kierkegaard envisions,” where the commitment is constantly re-affirmed by the agent, moment to moment. Yet, even given this response, it seems to me as though something like Fischer’s tracing example might explain the real difference at stake between these cases, rather than reasons-reactivity. I will discuss tracing in more

detail in Section 3.1., but for now we can consider that for our Kiergaardian<sup>64</sup> to have originally formed her commitment freely, the reasoning that led to it would have had to be reason-responsive (in both the receptive and reactive sense). At this stage, if the agent was incapable, even counterfactually, of ever not forming her Kiergaardian commitment, then perhaps it would be correct to say that she is not morally responsible. On the other hand, if, during this earlier process, there was a near possible world where she does not form the commitment for some relevant reason (or maybe she never reads Kierkegaard?), then this is a classic tracing example.

Distinct from possible responses available to Fischer and Ravizza, it appears to me that a person who will tell the truth even if the world would be destroyed, and *no possible reason could possibly sway them from their commitment*, is insane. Such a person might well be responsible for their actions in some sense, but not morally.

It is not clear to me if a commitment of the type Pereboom posits in his example is supposed to flow from a reason to action process (what I would call the System 2 oversight mechanism, what Fischer simply calls the actual-sequence mechanism) that is unreactive or from some motivation (pro-attitude) so powerful that the agent cannot possibly reason against it. If the former, then I would argue that we should hold the agent to be morally responsible *and* blameworthy, whereas in the second (which seems to me no different from an overpowering addiction or a severe psychological disorder – making this case similar to that of the psychopath discussed in Chapter 2: Section 1.1.) they would find the agent to not be morally responsible. Not seeking to conflate criminal and moral responsibility, but it is worth noting – as Schlosser does (2013: 224) – that:

[a] good indication that this is in line with the conditions on criminal responsibility is provided by the Model Penal Code's section on “mental disease or defect excluding responsibility” (§ 4.01). This section states that an agent is not responsible for criminal conduct if the agent “lacks substantial capacity either to appreciate the criminality [wrongfulness] of his conduct or to conform his conduct to the requirements of the law” (§ 4.01). Although this

---

<sup>64</sup> It may be worth noting that Kierkegaard himself would have certainly thought that in order to make the kind of moment to moment commitment of which Pereboom speaks, the agent must be making this commitment freely, with the real possibility of choosing against the commitment.

passage refers to “conduct as a result of mental disease or defect”, it clearly highlights the underlying presupposition that *one must be able to understand the reasons why certain acts are wrong and to adjust one’s behavior accordingly* [my emphasis].

In addition to reason-responsiveness, Fischer and Ravizza also laid out an *ownership condition* on guidance control (Sartorio has no comparable condition in her account). The idea advanced by Fischer and Ravizza is that for an agent to have appropriate ownership of the actual-sequence mechanism, it must be the case that she “*takes responsibility* for the mechanisms giving rise to her actions” (McKenna, 2009). Bratman (2000: 454) breaks down the three ingredients of what it is to *take responsibility* for a mechanism, according to Fischer and Ravizza, as follows: (I) one must see oneself as the source of one’s behaviour, in the sense that one sees that our pro-attitudes, beliefs and intentions result in changes in the world, (II) one must consider oneself a legitimate target for reactive attitudes based on how one employs the agency described in (I), and (III) the views held in (I) and (II) must be based on evidence. In other words, an agent can only be said to have guidance control if that agent meets two different subjective requirements, to be exact: a belief (backed by evidence) that she is the source of her behaviour, and a belief (backed by evidence) that she is a legitimate target for reactive attitudes. Additionally, Fischer and Ravizza argue that ownership in the form of “*taking responsibility*” is a “historical” notion (Bratman, 2000: 454), as whether or not an agent can be said to meet the requirements for ownership depends upon a certain history. This is important, as it aims to rule out the seemingly counterintuitive example where an agent is held to be morally responsible for a given actual-sequence mechanism that, while still reasons-responsive, was historically manipulated.<sup>65</sup>

In defending the historical character of the ownership of guidance control, Fischer and Ravizza utilise two types of examples: “tracing examples” and “manipulation cases”

---

<sup>65</sup> Consider, a case where an agent is, at some stage in her life, exposed to subliminal messaging that results in a powerful (but not irresistible) desire to kick her cat whenever she hears the word “Frankfurt.” In this case, the agent is still reasons-responsive and seems to meet requirements (I), (II) and (III), but according to Fischer and Ravizza, she is not morally responsible, as the actual-sequence mechanism does not meet the historical requirements for ownership. Their argument is that since the agent is unaware of the actual mechanism at work in this example, she cannot take responsibility for this mechanism.

(Bratman, 2000: 455). In tracing examples the agent acts out of a non-reasons-responsive mechanism, yet is still morally responsible because “the non-responsiveness of the mechanism is a predictable consequence of an earlier action that flowed from a reasons-responsive mechanism.” This would help in cases such as certain examples of drunk driving, where the moral responsibility carries over, so to speak, from a preceding reasons-responsive decision.

The second type of example – manipulation cases – is trickier; here the goal is to capture the intuitive idea that if an agent’s actual-sequence mechanism is manipulated in a suitable way (such as certain cases of hypnosis, brainwashing, subliminal advertising and direct stimulation of the brain) then the agent does not “own” the results in the appropriate way for moral responsibility. To make sense of this intuition within the frame of their argument, Fischer and Ravizza assert that the agent cannot take responsibility for the actual-sequence mechanism at work, as this mechanism is epistemically opaque to her. As such, she does not have ownership of it, so she does not qualify as having guidance control. Although they do not explicate it as such, I think we can describe this historical requirement in terms of a specification of the content of (III), let us call it (III\*): in order to take responsibility for an actual-sequence mechanism, the evidence an agent uses to justify (I) and (II) must include awareness of the actual history of that actual-sequence mechanism. Thus, even though they explicitly set out to provide an account of the volitional condition while setting aside the epistemic one, the ownership condition introduces certain epistemic demands that the agent must meet to have guidance control.

This aspect of Fischer and Ravizza’s argument – the ownership requirements for guidance control – is probably the aspect that has met with the most criticism. Alfred Mele (2006) argues that the subjective requirement makes the door for moral responsibility too narrow. It seems unproblematic to assume that in at least some cases an agent could be morally responsible even though she does not see herself as a legitimate target for reactive attitudes. Mele expresses this criticism using an example that runs like this: Imagine Sarah. Sarah is a committed hard incompatibilist, and so does not see herself as a legitimate target for reactive attitudes. Sarah, like most people, sometimes engages in telling lies to improve her own situation. Although it seems clear that we would find Sarah to be morally blameworthy for her lying,

since Sarah does not believe herself to be a legitimate target for reactive attitudes, it appears that Fischer would have to conclude that Sarah is not morally responsible for these lies. Similar examples, where one or the other of the subjective requirements is absent and leading to seemingly counterintuitive results, are not difficult to imagine.

Like Mele, Pereboom (2007) advances a criticism using an example: Imagine Susan. Susan has been created by neuroscientists such that they are able to modify her reasoning so that she reasons egoistically when they flip a switch. She remains reasons-responsive, and she is unaware of this manipulation and so holds the necessary beliefs for ownership. The manipulation is of such an art that even knowledge of the causal history of her actual-sequence mechanism (which is not itself manipulated by the neuroscientists) would not provide her with evidence to relinquish these beliefs. If her now egoist reasoning causes her to ignore a moral reason in favour of self-gain, would she be morally responsible? Fischer (2004: 158) says “yes,” with the important caveat that even if Susan is morally responsible in this case, she is not blameworthy. Using my terminology, Fischer is maintaining that Sarah is open to moral responsibility, but should not be held morally responsible. What makes this so implausible is that usually when an agent is open to moral responsibility for an action with moral valence, the agent *can* be held responsible. Indeed, this seems like a necessary relationship. That said, Fischer could potentially defend his position by arguing that although moral blame may not be legitimated, the reactive attitudes associated with what Williams (1976) calls agent-regret may be justified, which softens the implausibility. That said, this remains a shortcoming that his approach struggles to deal with in a wholly satisfactory fashion.

Michael Bratman (2000: 454-458), despite agreeing that “[m]oral responsibility may be globally historical,” contends that the arguments presented by Fischer and Ravizza are insufficient to show that this is the case. He argues that the tracing examples only show that *sometimes*, in *some* cases, the history of behaviour can play an important role in determining moral responsibility. This is insufficient to show that the history is a *necessary* consideration in regard to moral responsibility. On the other hand, he argues that Fischer and Ravizza’s lack of exposition regarding the exact mechanism

by which reasons lead to actions means that their treatment of manipulation cases falls well short of proving what they claim it proves.<sup>66</sup>

Overall then compatibilists are best off arguing that it is the reason-responsive component of guidance control (or some similar AS replacement thereof), without the addition of an ownership condition, that is necessary for moral responsibility, and that this type of control is compatible with the truth of determinism. Indeed, the influence of this understanding of the control condition might be best evinced from Pereboom's (2007: 199) admission, despite his position as a hard incompatibilist and his criticisms of the ownership requirement, that:

I suspect I can agree with Fischer on the following claim: his theory of moral responsibility – guidance control spelled out in terms of reasons-responsiveness – provides the most promising account of what might be the most significant sense of moral responsibility that can be retained given the best philosophical arguments and the best scientific theories we have about the physical world.

For her part, Sartorio (2016) proffers an AS replacement for Fischer and Ravizza's understanding of reason-responsiveness. She argues that their understanding fails to capture the crucial insight of the AS model: that only facts pertaining to the actual sequence should be relevant to determining the presence of control, what she calls the "Exclusiveness thought" (*ibid.*: 113). She contends that their view involves the consideration of facts concerning merely possible worlds, rather than exclusively those pertaining to the actual sequence, since in their account it is necessary to consider certain modal properties of the actual-sequence mechanism. In contrast, on her account the absence of certain reasons in of themselves reflects an agent's sensitivity to reasons, without any need to make claims about possible worlds. By absence of a reason here is meant both that the reason was not present, or that the agent does not take there to be such a reason. So, when an agent fails to take some consideration as a reason to act, this "absence" of a reason explains the agent's behaviour, and reflects the agent's sensitivity to reasons, as much as the reasons to which the agent did in fact respond. I consider both Fischer and Ravizza's, and

---

<sup>66</sup> To appreciate the full scope of these challenges to Fischer's arguments, see Bratman, 2000; Pereboom, 2007; Mele, 2006.

Sartorio's accounts to be live options, and will not adjudicate between them in this dissertation.

As mentioned already, Fischer and Ravizza take their account of guidance control to explain only the volitional condition, explicitly setting the epistemic condition aside. Starting from the opposite direction, many volitionist accounts begin with the very compelling claim that “morally responsible agency is (directly or indirectly) conscious agency” (Levy, 2008: 214). The defenders of this claim are keen to point out that only that behaviour that results from conscious control can be open to moral responsibility, since:

Consciousness serves the function of allowing parts of the brain that are otherwise relatively isolated from each other to communicate... The global workspace [of conscious thought] allows all the mechanisms constitutive of the agent, personal and subpersonal, conscious and unconscious, to contribute to the process of decision-making. Hence conscious deliberation is properly reflective of the entire person, including her consciously endorsed values. (*ibid.*: 220)

This position has an obvious appeal, after all, we ourselves feel most immediately responsible for our conscious choices. It can also be argued that such consciousness is necessary to meet the epistemic condition for control: i.e. for an agent to know what they are doing; the agent must be consciously aware of doing so. Levy (2011: 114-116) pushes this further, arguing that not only must the agent be consciously aware of the behaviour in question, but “must properly *appreciate* the significance of bringing about that state of affairs, where the significance of a state of affairs consists of the features which provide reasons for bringing it about (often, but not always, moral reasons).” On a view of this sort, moral blameworthiness would seem to be limited to akratic behaviour, and moral ignorance would always exculpate.

Yet as we say in Section 2, it certainly does not seem the case that we limit the behaviours for which we hold ourselves and others blameworthy to those that are under *conscious* control. To illustrate this, consider these three examples:

*Unthinking rescuer:* Owen is standing in a subway station waiting for his train. Suddenly, another man falls onto the tracks below. Without hesitation, and without deliberation or thought to the consequences, Owen jumps down and pushes the other man flat between the tracks as the train passes over them both. By doing this Owen saves the other man's life. When asked afterwards about his actions, Owen says, "I didn't even realise I had jumped down until I had done it, not consciously anyway."

*Forgetting:* Mike has driven to the store with his infant child in a baby chair on the back seat. Having arrived at the store, Mike leaves his child in the vehicle, planning to only step out for a few minutes. In the store, Mike runs into an old friend who begins to chat with him, and Mike forgets about his child in the car. After a long conversation, Mike makes his way back to the car to find that the child has sadly died of heat exhaustion. When asked about these actions Mike replies, "I forgot about him, it entirely slipped my mind."

*Unconscious bias:* Greg is a lecturer at a university, and is busy grading papers. Greg has an unconscious bias against female students, tending to award them lower grades than their comparable male peers. When confronted about this discrepancy by a female student, and consequently made aware of his bias, Greg says, "I hadn't realised that I had been influenced by this unconscious bias."

In all three cases it seems reasonable to say that we would find all three these agents to be morally responsible for their actions: Owen being praiseworthy, while Mike and Greg are blameworthy. There may be some doubts about the case of Mike, but even if this example is not compelling the cases of Owen and Greg surely are. This would appear to be so even though in none of these cases does the agent in question have conscious control over the morally pertinent behaviour, because in none of them does the agent have sufficient epistemic access to what they're doing and/or the reasons they are doing it for or should not be doing it for. For Levy, this leads to the conclusion that none of these agents should be held morally responsible (disregarding for the moment the possibility of *tracing*, which will be discussed shortly), but this is a radical revision of our responsibility practices, and though Levy and others<sup>67</sup> may be prepared to accept such a revision, I argue it should be resisted if possible. And

---

<sup>67</sup> Such as Pereboom (2007; 2013) and G. Strawson (1994) for example.

attributivist accounts claim to offer an alternative account of moral responsibility that can justify such resistance, as we will see in the next section.

There are attempts by volitionists to put forward less revisionary responses to the types of cases presented here, and such attempts frequently make use of the notion of tracing as introduced in the discussion of guidance control. The idea here is that an agent could be open to responsibility for some given behaviour even though that behaviour was not under her control (usually, but not necessarily, due to a failure to meet the epistemic condition) provided that the behaviour in question can be shown to originate from some foregoing behaviour – what Smith (1983) calls the *benighting* act, though in this case act could include an omission or effect – that was under appropriate control. The paradigmatic example of this is that of the drunk driver: Drinker is drinking at a bar. He gets so drunk that he is no longer able to respond to the moral reasons not to drive while drunk. He gets in his car and drives, resulting in a collision with another vehicle that is Drinker's fault. In this case, though Drinker did not have conscious control over his conduct, he is still open to responsibility as the causing of the collision can be traced back to a point in the past where Drinker had a moral reason not to keep drinking without planning for his trip home, but continued to drink. This would then be the benighting act, as Drinker had conscious control over it. However, this strategy of tracing has serious limitations. It does not seem that all, or indeed many, of the sorts of non-consciously controlled behaviours for which we seek to hold agents responsible can be traced back in this way. There is also the additional problem that such tracing explanations can be thought to misidentify the locus of responsibility. In other words, it seems wrong to say that Drinker is morally responsible *only for his decision to keep drinking*, rather we want to say that he is responsible *for the collision he caused*. We have not heard the last of tracing, as it plays a role (though a limited one) in my own account, but for now we move on to consider if attributivist accounts can provide a more compelling approach to understanding moral responsibility.

### 3.2. Attributivism

Given the limitations of volitionism, it should be no surprise that non-volitionist accounts have risen as competitors. The most influential such competing account is

that of attributivism, which comes in several flavours: character based accounts, real or deep self accounts, and Quality of Will (QoW) accounts. As was explained earlier, what unites this wide variety of approaches is that they all contend that a given behaviour can only be the basis for an agent being open to responsibility if that behaviour is genuinely revelatory of the agent who performs it. The different approaches are differentiated by how they answer the question: *what exactly must be revealed?* Character-based accounts (often called Neo-Humean accounts), as might be readily guessed, argue that what must be revealed is some aspect of the agent's character. Real self accounts differ in that what must be revealed is taken to either be the agent's Appetites (Frankfurt proposes a seminal example of this approach), or the agent's Reason (which is the approach favoured by Watson). Finally, QoW accounts argue that what must be revealed is the quality of the agent's will, which is usually understood in terms of the degree and quality of care exhibited by the agent. It is the last of these approaches to which I will be giving most of my attention – as they are, to my mind, the most compelling as well as the most pervasive in the recent literature – but before this it is worth considering some shortcomings of the alternatives.

Straightforward character-based versions of attributivism will always have the difficulty of accounting for “out of character” behaviour, examples of which seem to be plentiful. A man may be miserly and close-fisted, but in one extraordinary – and perhaps personally mystifying – instance he gives out money to pay for the food, bed, and board of a destitute and desperate person. It seems that the man should be considered praiseworthy for this, even though the action does not in fact reveal his character, which is in fact niggardly. The same is true (even more vividly so) when considering blame. We could imagine an agent with a morally upright character who, in what is again an extraordinary instance where he himself might struggle to reconcile it within himself, performs some morally bad action – say, kicking his dog for no reason, or brutally mocking a co-worker in front of others. These out of character moments might often have explanations: the miserly man might have recently heard that his grandson was born, and unbeknownst to him been moved by this to act in an uncharacteristic fashion, and the morally upright man might have had a night of terrible insomnia that frayed his temper. Now these explanations might move us to reduce the degree of responsibility present, or to *excuse* the behaviour, but it seems implausible to think that the agents are not responsible at all. However, if we

hold that the actions for which an agent can be held morally responsible are only those that reveal the agent's actual character, then these actions would simply not fit the bill.

Moving away from talk of character, or at least employing a different understanding of what the relevant notion of character should be, so-called “real self” or “deep self” accounts<sup>68</sup> contend that what an agent’s behaviour must reveal is one of either the agent’s *Appetites* or *Reason*. By Appetites is meant the agent’s desires, and usually the focus has been on the agent’s higher-order desires (Frankfurt, 1988). In contrast, the agent’s Reason is identified with her evaluative judgements (Scanlon, 2008), beliefs, or valuational system (Watson, 2004). In simple terms, both Appetite and Reason form important facets of the character of an agent, ones that can be revealed by an agent’s behaviour (and attitudes, though I do not discuss these), but different real self accounts disagree on which of the two reflects the genuine agent. As it turns out, either path leads to serious objections: Appetite views struggle in accounting for responsibility in cases where an agent acts against her desires, up to and including her higher-order ones. This is most apparent in cases where the agent acts in a praiseworthy fashion against her desires, such as in the example put forward by Arpaly and Schroeder (1999: 171) of “a person who commits a brave moral act despite continuous desire for escape.” Similarly, Reason views struggle with responsibility in cases of akrasia and inverse akrasia. By inverse akrasia is meant a case where an agent acts in line with the balance of reasons, but takes herself to be acting against it – the classic example being Hursthouse’s (1991) discussion of Huck Finn. In these cases, the agent’s actions are not reflecting their valuational systems or evaluative judgements, or their beliefs about what they should do, and so on a Reason view of the real self approach, they could not be morally responsible for them.

In the face of these shortcomings, some have attempted to provide more integrated notions of the real self, such as the “whole self” account put forward by Arpaly and Schroeder (1999). On this account,

---

<sup>68</sup> Some use the term “attributivist” to refer specifically and only to proponents of the real self approach. Though I recognize this usage, it will not be what I mean by the term in this work. As mentioned before, I will be using the term to refer to a family of non-volitionists approaches.

other things being equal, an agent is more praiseworthy for a good action, or more blameworthy for a bad action, the more the morally relevant psychological factors underlying it are integrated with her overall personality. (ibid.: 172)

The psychological factors in question here are *desires* and *beliefs*. And for them to be well-integrated they must be (i) deep, and (ii) not in opposition to other deeply held beliefs or desires. Depth is determined by the degree to which a psychological factor determines an agent's behaviour. Beliefs oppose each other when they cannot be simultaneously true, desires when they cannot be simultaneously satisfied

This combined model does give Arpaly and Schroeder leeway in tackling the problems that confront real self accounts that focus solely on Appetite or Reason, though it is not without objections.<sup>69</sup> However, in their more recent book, *In Praise of Desire*, they have developed a more clearly desire-centered account of responsibility. According to this account, what must be revealed in order for an agent to be blameworthy or praiseworthy are facts about her intrinsic desires. To be blameworthy for an action is “to act out of ill will (an intrinsic desire for the wrong or bad) or indifference to the lure of the right or good,” (2013: 159) and vice versa in the case of praiseworthiness. They argue that acting for moral reasons necessarily involves acting from an intrinsic desire for the right or the good, and that it is behaviour caused by intentions backed by such reasons that form the basis for moral responsibility (2013: 87). There remain difficulties here, particularly with how Arpaly and Schroeder deal with addiction and habits. In attempting to explain why addiction should serve as an excuse on their account – a challenge since an addict seems like a paradigmatic example of an agent acting from an intrinsic desire – they claim that the actions of the addict stems from habit rather than from desire. This is a problematic response, as it is unclear, and rather implausible, that actions that spring from habit are inherently less blameworthy than those that spring from desire – indeed, an agent for whom cruelty has become a habit seems in some ways *more* blameworthy (Holton, 2015).

Criticisms aside, given the talk of good and ill will in Arpaly and Schroeder's account, I am inclined to consider it as an example of the QoW approach, to which I

---

<sup>69</sup> See Shoemaker (2015b) for a strong critique of the Whole Self view.

now turn. Another compelling example of a QoW account – and the one I take to be the most promising attributivist account overall – is the *explanatory quality of will account* recently developed by Björnsson (2017a; 2017b). Central to this account, he argues for the following condition on moral responsibility (2017a: 148):

MORAL EXPLANATORY BLAME (CREDIT): X deserves moral blame (credit) for Y if and only if Y is morally bad (good) and is explained in a normal way by X's quality of will falling below (above) what could be properly morally demanded of X

By *quality of will* here he means (ibid.: 149):

[H]ow well she cares about what is morally important, where caring about something in the relevant sense involves being disposed to pay particular attention to information relevant for promoting or not obstructing the object of caring and to have one's behavior be guided by such information.<sup>70</sup>

And the blame and praise (or credit) involved can come by degrees (2017a: 148):

DEGREE FROM DEVIATION: The degree of blame (credit) X deserves for Y depends on the value of the outcome and on how much of a deviation of the agential aspect from what can be properly demanded is required in the normal explanation of the object of blame (credit).

---

<sup>70</sup> This idea of caring for what is morally important is very similar to Arpaly and Schroeder's notion of the good understood in terms of intrinsic desires. To care about what is morally important could be understood as having an intrinsic desire for the right or good. Yet there is a noteworthy difference, that being that for the latter pair of thinkers a good will can also be a case of indifference to the temptation of the intrinsic desire for the wrong or bad, and ill will can be the result not only of an indifference to the intrinsic desire for the good, but also an intrinsic desire for the wrong or bad. I am inclined to agree with Björnsson here. We can imagine an agent who is *incapable* of possessing the intrinsic desire for the right or good – we might say, incapable of care for what is morally important – but who does possess the intrinsic desire for the wrong or bad. Such a being, let's call it Demon, would not seem to be an apt target for moral responsibility. Our moral responses would be meaningless in this case. Our response to such an entity would presumably be to treat them as we might treat a dangerously insane person. In a similar vein, to be indifferent to the intrinsic desire for the wrong or bad does not seem praiseworthy unless the reason for this indifference stems from the right source. It seems more accurate to say that acting from good will means acting from the intrinsic desire for the right or good, and acting from ill will means acting with insufficient responsiveness to this desire.

A central concern of Björnsson in developing this account is that it must explain under what conditions a lack of awareness excuses an agent from responsibility, as well as how to understand the fact that in our responsibility practices we often hold agents responsible for failing to respond to considerations that they *should have known*. He is most keen to reject the position of certain volitionist accounts that demand such a stringent epistemic condition on moral responsibility that such responsibility would be limited to akratic behaviours. However, he goes further than this rejection. He argues that there is *no need* for either a separate volitional or epistemic condition – as volitionists might claim<sup>71</sup> – once his explanatory quality of will condition is suitably understood. To arrive at such a suitable understanding, two crucial elements need to be expanded upon: what is meant by the phrase, “explained in a normal way”? And what determines the degree of care that can be “properly demanded” of an agent.

To the first of these we are given the following explanation (2017b: 150):

In addition, the explanation has to be *normal* in some relevant sense. Suppose that I am just about to finish a decent kitchen cabinet in spite of my poor craftsmanship, when some spiteful person secretly changes the dimensions of some of the parts because they think that someone with my laughable skills doesn’t deserve a nice cabinet. Then the end result is a bad cabinet, and it is bad in part because of my poor craftsmanship — it is what triggered the sabotage. Still, since the explanation is abnormal, I am not to blame for the cabinet’s sorry state.

Björnsson (2017b: note 6) also mentions that what makes an explanation a normal one is not merely likelihood, and that the cases that should be taken as relevant must be those that the practice in question is designed to track. However, he also recognizes that this is less than a complete account of the normalcy requirement. Despite this, if we take a more or less intuitive sense of normalcy (like that pointed to by the example

---

<sup>71</sup> As with the role of control generally, it should not be thought that attributivist accounts (such as those of Watson, Scanlon or Arpaly and Schroeder for example) reject that epistemic considerations do sometimes play a role in determining responsibility. However, they do reject that there is a necessary and determinate epistemic condition on moral responsibility. Rather, lack of awareness or appropriate belief only matters when this breaks the link between an agent’s behaviour and the morally relevant feature of the agent. In this regard, Björnsson’s account is in line with the general attributivist strategy.

of the kitchen cabinet-maker), we can see that it would not be abnormal to think that, at least sometimes, an agent's degree of care can explain some morally valent behaviour, X, without the agent having conscious control over X, or perhaps even having the unaccessed belief that she is X'ing. However, as we will see in Section 4.3, I argue that the explanatory quality of will account can be improved by developing a more detailed understanding of the notion of normalcy at work, and combining this with an endorsement of pluralism about moral responsibility.

The second element, determining what degree of care can be properly demanded of an agent, is directly answered by the provision of the following principle (2017a: 149):

DEMAND FROM CAPACITY: The degree to which it can be properly demanded that X cares about something depends only on X's capacity for so caring, not on how X came to care to the degree she does

DEMAND FROM CAPACITY indicates a limit to the question of both how historical and how deep the explanatory quality of will account goes. I take this to be a virtue of the account, as it helps the account capture responsibility in out-of-character cases and so-called manipulation cases without having to posit claims about the history of the agent or the agent's actual-sequence mechanism – contra Fischer and Ravizza. So, for example, an agent who has been manipulated by neuroscientists into having a reduced capacity for care, would not be expected to demonstrate the same degree of care as she would have had the manipulation not taken place. And this is no different to the case in which this reduction was the result of something less out of the ordinary, such as poor formative circumstances or substance use. This is important, since it means that whether or not the agent in question was aware of the reduction to their capacity to care, and aware of the history behind it, is irrelevant for determining if this capacity is present. At the same time, reductions in the capacity to care will reduce the openness of the agent to blame, both in terms of the types of cases where she will count as blameworthy, and more usually the degree to which she will be blameworthy. It also has an advantage over Arpaly and Schroeder's desire QoW account since it allows the explanatory quality of will account to make sense of responsibility in cases of addiction, as in such cases it is plausible to say that the agent's capacity for caring is impaired. This would explain why the presence of

addiction of a certain sort seems to diminish the degree of responsibility an agent might have for a behaviour, or indeed, if the case is extreme enough, be exculpatory.

Turning back to the epistemic condition, it should be clear that though at times a lack of awareness or belief may break the explanatory link between behaviour and quality of care, there are many cases where it will not. And it seems plausible to think that those cases where it does not are precisely those cases where we would still want to be able to legitimately attribute responsibility, such as in *Unthinking rescuer*, *Forgetting*, and *Unconscious bias*. Admittedly, much hinges on exactly how the notion of normalcy at work in MORAL EXPLANATORY BLAME (CREDIT) is fleshed out, but overall I take this account to be the strongest attributivist option.

### 3.3. Concluding remarks

There are three important conclusions from this section. The first is an observation, and the remaining two are my judgements as to the most compelling contender from each of the two approaches discussed.

The observation in question is that *all* the accounts of responsibility considered here, volitionist and attributivist (of all stripes), begin with the assumption that for an agent to be open to responsibility on the basis of some behaviour requires that this behaviour *say something* about the agent in question. The accounts differ on what this something should be: for the volitionists it tends to be volitional and rational agency, whereas for the attributivists it has been taken to be character, Appetite, Reason or care (which may or may not be best understood as an intrinsic desire).

In terms of volitionist accounts, I take the volitional condition on moral responsibility to be best understood as reason-responsiveness, whereas a compelling account of the epistemic condition remains elusive. Neither the demand for conscious awareness endorsed by Levy, nor the ownership condition introduced by Fischer and Ravizza – which though not explicitly an epistemic condition does smuggle in requirements about what an agent must know to be open to responsibility – are promising candidates. And examples such as *Unthinking rescuer*, *Forgetting*, and *Unconscious bias* seem to be deeply problematic for volitionist accounts because of this. So,

although there remains something persuasive in the idea that some form of control is important for moral responsibility, it seems that the volitionist claim that this relationship is necessary has an uphill challenge before it.

When it comes to attributivist accounts, character-based and real self accounts struggle to deal with certain counterexamples, and given that the allure of attributivism is not insignificantly tied to its ability to capture a wider gamut of our responsibility practices than volitionism, counterexamples such as these cannot be easily dismissed. QoW accounts, on the other hand (at least those of Arpaly and Schroeder, and Björnsson), take the agent's *moral concern* directly – understood either as her intrinsic desire for the right or good (or lack thereof), or as the degree of care for what is morally important – as that which must be revealed in, or explanatory of, a behaviour in order for an agent to be morally responsible for it. Though either approach is in a better position to deal with the counterexamples than the accounts that went before, Björnsson's explanatory account has the advantage of more plausibly explaining cases where the presence of addiction seems to reduce an agent's fittingness for moral responsibility, thanks to DEMAND FROM CAPACITY. For these reasons the explanatory quality of will account strikes me as the best contender amongst the attributivist accounts, and I will be borrowing substantially from it in developing my own account of responsibility in Section 4.3. and 5.

#### **4. Ambivalence and pluralism**

In this section I will present what I take to be a serious objection to both volitionist and attributivist accounts of moral responsibility, and then propose a potential solution. The objection in question is that neither of these types of accounts can properly capture those cases of moral responsibility where we experience an abiding and deep ambivalence about the kind of moral response that the given case calls for. To explain this objection, I first describe two important assumptions that pervade many accounts of moral responsibility: invariantism and conservatism. The first of these holds that the conditions for moral responsibility should be invariant, or exceptionless, while the second acts as a criterion for choosing between theories of moral responsibility by asserting that a theory of moral responsibility is more

convincing than another, all else equal, if the former captures more of our responsibility practices.

Having introduced these two assumptions, I then present an argument for why they cannot be mutually held, and furthermore, that invariantism about moral responsibility should be rejected. The argument I present leans on one already eloquently made by David Shoemaker (2015a): that there are some cases of moral responsibility where our experience is not one of indecision regarding whether moral responsibility is present, but rather ambivalence about which of our moral responses seem to be legitimated by the responsibility present. I will argue that no invariantist account of moral responsibility can hope to capture this ambivalence, and that capturing it would strengthen an account of moral responsibility in line with conservatism. As such, we should adopt a variantist or pluralist understanding of moral responsibility, as this is the best match for our actual responsibility practices. In particular, I will defend a dual model of moral responsibility based somewhat on that put forward by Watson (2004). On this model, there are two distinct types or variants of moral responsibility: *attributability* and *accountability*. Despite this, and contra Shoemaker, I will argue that this pluralism should not be understood disjunctively, as any case where moral accountability is present, moral attributability must also be present, though not vice versa.

I will conclude by contending that accepting this pluralism allows us to see that in many cases the dispute between volitionists and attributivists rests on an equivocation about what is meant by *an agent being open to moral responsibility*. Where volitionists take the moral responsibility at stake to be accountability, attributivists are primarily concerned with attributability.

#### *4.1. Invariantism and Conservatism*

As I have discussed, there have been many and varied accounts of moral responsibility put forward over the years. However, regardless of the interpretation of moral responsibility being advanced, for most of this history it was almost universally presumed that the concept of moral responsibility is *unequivocal* and *unitary*. In other words, it was presumed that there is only a single “variant” of moral responsibility,

and that this single version covers all cases that we would refer to as cases of moral responsibility, cases where an agent is a legitimate target for moral praise or moral blame. This is obviously not meant to imply that all agents that are morally responsible are equally blameworthy or praiseworthy. On almost all accounts, some actions will presumably be deemed praiseworthy while others are counted as blameworthy, and both praise and blame will come by *degrees*. Regardless, these differences are not considered to be the result of differences in what it means to hold an agent morally responsible. Rather, the difference lies in the fact that what they are morally responsible for has a different moral valence.

The most common form this kind of thinking takes runs (loosely) as follows: it is assumed that there are some moral principles of right action. When an agent violates one or more of these principles, she is potentially morally responsible for this. Presuming that shoplifting and murder both violate accepted moral principles of right action, an agent who shoplifts and an agent who murders are both morally responsible for their actions. We may well treat these agents differently – the murderer may be a legitimate target for more severe responses than the shoplifter for example, as we think of the murder as *more* blameworthy – but this difference is not because the agents are morally responsible in different ways. This is the case regardless of whether we adopt a volitionist or attributivist stance. If we adopt a volitionist view, then the degree of blameworthiness or praiseworthiness will depend on the moral valence of the behaviour in question, as well as the agent's degree of control over it. It seems uncontroversial to assume that in most cases (if not all) where the two agents have equal control we will rule that the murderer will *deserve* more blame than the shoplifter, though if the murderer was suffering from a severe psychological condition that undermines control it is possible that this verdict may alter, depending on the question of degrees. On the other hand, if we adopt an attributivist view, then although the moral valence of the behaviour will still play the same role, the appropriate reaction will in this case be determined by the *extent* to which the given behaviour reveals whatever morally relevant feature of the agent the particular attributivist account we employ puts forward, and the *quality* of this agential feature. The recognition that moral blame and praise comes by degrees is proper and necessary for any compelling account of moral responsibility, given how this is a central feature of our responsibility practices, however, as we will see, there is a

problem that emerges from the endorsement of an understanding of moral responsibility that comes *only* by degrees.

Turning back to the assumption that moral responsibility is unitary: even in accounts of moral responsibility that do not assume any pre-given moral principles of right action – such as Strawson’s (1962) influential account – moral responsibility is still treated as unambiguously unitary. To be held morally responsible always means the same thing, though there are differences in degree, and in the results of being held so responsible. Part of the motivation for this way of thinking about moral responsibility is not difficult to surmise. In everyday language, after all, we do not talk about different types of moral responsibility, though we do talk about degree. Indeed, it often seems that the question of whether an agent is morally responsible for a given activity or its outcomes can be decided in complete independence from the question of what would follow from this responsibility. However, even on the consequentialist view, on which an agent is only to be held morally responsible when doing so brings about desired changes in the agent or the agent’s behaviour (or in the behaviour of other agents), and so one cannot determine whether an agent is a legitimate target for moral responsibility without considering the consequences, it remains the case that the conditions for moral responsibility can be formulated as an invariant principle.

In contrast to this longstanding presumption, there is a trend in a considerable amount of the recent work surrounding moral responsibility to move away from a view that takes the discussion to be concerned with a single, unitary, and invariant concept of moral responsibility, and towards one that takes the discussion to involve a number of different, but related, concepts of responsibility. This movement has found its contemporary impetus in the work of Gary Watson (2004), with David Shoemaker (2015a) as perhaps its most sophisticated proponent. Adopting this position seems to hold the promise of helping to dissolve some of the longstanding oppositions in the discussion – such as between volitionists and attributivists – by revealing that the opposing accounts simply have different concepts of moral responsibility as their objects, and that the disagreement arises from a failure to distinguish between these different, but related, concepts. Despite the facts of everyday language use mentioned above, these more recent discussions around moral responsibility have often focused on the question of whether the concept of moral responsibility is indeed unitary as the

traditional views take it to be. There seems to be a growing position that there may be several, co-extant and equally essential, variants of moral responsibility.<sup>72</sup>

To understand the structure of the argument supporting this move, it is useful to start with the insights outlined in Doris, Knobe, and Woolfolk's 2007 paper, "Variantism About Responsibility". In this work, the authors identify what they call "two dogmas of responsibility" (2007: 183), namely: *invariantism* and *conservatism*. They then proceed to argue that, given experimental evidence drawn from a number of survey experiments, both these dogmas cannot be simultaneously maintained. An *invariantist* account is one that posits "*exceptionlessly relevant* criteria for responsibility attribution" and requires that the same criteria be employed in any given case (2007: 184). Most accounts of responsibility endorse invariantism in this sense, though it should be noted that invariant criteria can be very complex, vague, or posit "paradigms or prototypes" as opposed to necessary and sufficient conditions. However, this added nuance does not alter the reality that these accounts aim at the provision of some exceptionless principle that will serve to govern our attribution of moral responsibility. For illustration of the widespread nature of the endorsement of this "dogma," Doris et al. presents the necessary conditions for moral responsibility that have been advanced by a number of different thinkers, in order to display their tacit assumption of invariantism. These include the common conditions of source and leeway incompatibilist accounts, and the compatibilist accounts of Frankfurt (1988), Wolf (1990), and Fischer and Ravizza (1998).<sup>73</sup> Not wishing to reproduce every example given in their argument, I will present only three here. In the case of source incompatibilism this condition is that the agent must be the ultimate source of her behaviour – the "ultimacy" condition – whereas in leeway incompatibilism it is the presence of alternate possibilities (Kane, 2002: 2007, and Haji, 2002). As an example from the compatibilist position, they present Frankfurt's condition that the responsible agent must suitably identify with her behaviour and motives in the form of their higher-order attitudes. To add to this list, even free will skeptic accounts, such as that of Galen Strawson (1994), still posit invariant conditions for moral responsibility,

---

<sup>72</sup> See Watson, 2004, Darwall, 2006, McKenna, 2012, Arpaly and Schroeder, 2013, Scanlon, 2015, and Shoemaker, 2015a amongst others.

<sup>73</sup> It is worth noting that Fischer and Ravizza do in fact recognize that there may be other variants of moral responsibility, and so are not truly invariantist as Doris *et al.* takes them to be.

with the important caveat that they simply maintain that these conditions are never actually met.

By *conservatism* is meant the view that “folk belief is a constraint on philosophical theorizing” (Doris et al., 2007: 185). One way to understand this constraint is to point out that an account of moral responsibility that falls too far afield of the actual practices of moral responsibility for which it is meant to account runs the risk of answering the wrong question. After all, we are interested in how to understand the moral responsibility we have. Given the strong reliance on cases, and the intuitions they elicit, by thinkers on all sides of the discussion of moral responsibility, it seems clear that some presumption of conservatism is prevalent throughout the literature. Even when the thinkers of a given view argue against some set of practices, as hard determinists may do in the case of blaming or praising practices in their entirety, they often precede by starting from a folk response to some case, and then generalizing this response in order to show that it stands in contradiction to other responses and/or practices.<sup>74</sup> Doris et al. also note that a commitment to conservatism does not mean that the folk are assumed to be infallible. Rather the situation is more like that of Rawls’ “reflective equilibrium” (*ibid.*), as we seek a coherent balance between our particular judgements and our general principles. So though in some cases we can surely say that the folk are in error (which may demand an appropriate error theory), an account which conserves more of the folk responses than another, *ceteris paribus*, should be preferred.

Doris et al. further contend that the results of recent experimental philosophy problematize the possibility of maintaining both invariantism and conservatism. These experiments have generally taken a familiar form: a group of respondents are provided with a case, or set of cases, involving some potentially divisive elements as regards the responsibility present, and are then asked to provide judgements as to the moral responsibility of one or more agents involved in the case. The results of these surveys seemingly show that folk responses are strongly variantist. It does not seem likely that all the disparate responses of the folk could be explained by anything approaching a single reasonable principle (or a small set of principles). This leaves us with a tension between two unsavoury options, forego conservatism about the folk

---

<sup>74</sup> A good example of this strategy is Derk Pereboom’s Four-Case Argument (2014).

responses, or give up invariantism. However, I will *not* be employing these arguments from the experimental literature here, partly because they are – even if true and compelling – only able to necessarily motivate pluralism about the conditions for the *attribution* of moral responsibility, rather than pluralism about the concept of moral responsibility itself. Furthermore, I think that certain attributability accounts, such as Björnsson's, can in fact account for the folks' apparently invariant responses in these cases, once certain confounding factors are accounted for. For these reasons, though the interplay between invariantism and conservatism is important, I take it to be a different set of cases that should convince us to adopt pluralism about moral responsibility. These I discuss in the next section.

#### *4.2. The problem of ambivalence cases and pluralist responses*

The central pillar of support for pluralism about moral responsibility is founded on what Shoemaker (2015a: 1-4) describes as cases of moral responsibility that engender in us *deep ambivalence*. Shoemaker provides four examples of such cases, but I will only consider two of them, as these are the only two that deal with responsibility for *behaviours*:

Firstly, the example of Skip, who is a psychopath – lacking both empathy and conscience. He is charming, manipulative, and capable, rising to great heights in the company in which he works. Indeed, he makes himself so valuable to the company that they settle multiple lawsuits that are brought against him for sexual harassment and similar offenses, including one case where he broke a woman's arm when she resisted his attempt to force her onto his lap. He said of this woman, “She's insane. She broke her own arm. She struggled with me, the stupid bitch. Why the hell did she put up such a fight?” (Shoemaker, 2015a: 1). It is important to note of Skip that he is obtuse to calls for him to provide some moral accounting of his behaviour, as shown when his mother asks him shortly before his wedding to a billionaire's daughter Juliette, “why he had to marry her, why he had to do this to Juliette's life. Tempted to ignore her at first (as usual), Skip smiled and said, ‘We both know she'll never know what hit her.’”

Secondly (though the fourth example Shoemaker gives), is that of Robert Harris, a famous serial killer. Shoemaker is not the first thinker to discuss Harris' case, and he draws considerably on Watson's (2004: 234-242, 280-281) discussion thereof. Harris murdered two young men and then calmly ate the fast food that they had ordered. Further still, he "joked" to his brothers that they should pose as police officers and inform the boys' parents that their sons had been killed. Indeed, even after he was caught, "his loathsome personality was loathed by guards and even his fellow inmates" (*ibid.*: 3) And yet:

Harris's upbringing, however, was unimaginably terrible. His parents were alcoholics who repeatedly terrified, beat, and abused him. At 14, he was sentenced to a youth detention center, where he was regularly raped and beaten. He attempted to commit suicide twice. By the time he was 19, it is no wonder that he had begun killing and torturing animals. His adult criminal life had begun.

What is disconcerting in these cases is that it seems as though in some sense the agents can be held morally responsible: it seems right to think that Skip and Harris are blameworthy for their actions, certainly worthy of our disdain and contempt. Yet, it also seems as though in another sense they are not blameworthy: reactive attitudes such as angry remonstration or indignant railing seem inappropriate, and simply pointless in these cases. In Shoemaker's (2015a: 3) words, when confronted with these cases:

My own reaction, and the reaction of others who have written about such cases, is a profound unease. This is not, however, the unease of uncertainty. Rather, it is the unease of *ambivalence*. The reaction, in other words, is that these agents seem worthy of some responsibility responses but not others, which suggests that they are responsible in some ways but not in others.

Crucially, existing theories of moral responsibility that exhibit invariantism are incapable of properly accounting for this. On a volitionist account such as that of Fischer or Levy, it seems that Skip and Harris would simply be off the hook, as for the former Skip would not qualify as a moral agent, and Harris is likely to fail in

meeting the ownership condition – not seeing himself as an apt target for reactive attitudes – and for the latter both Skip and Harris are incapable of meeting the epistemic condition on moral responsibility. Yet this conclusion strongly violates conservatism. An attributivist account such a Björnsson’s might find better purchase here, as in both cases a dearth of care seems to be the grounds for our initially wanting to attribute responsibility (though it may be questioned whether either of these agents had sufficient *capacity* to care). However, the explanatory quality of will account only comes by degrees, not type, and so without expansion it cannot explain why in the case of Skip and Harris it seems that some of our moral responses are justified, but not others. I will argue that the explanatory quality of will account can in fact accommodate a solution to this problem, but it requires a *pluralist* account of how to fill in “explained in a normal way” in MORAL EXPLANATORY BLAME (CREDIT).

Shoemaker uses this failure of existing accounts to explain the ambivalence cases as support for his proposal of a *tripartite theory of responsibility* (2015a: 16), according to which there are three types of responsibility, with each conditioned by a different facet of the quality of an agent’s will: *attributability*, which is conditioned by the agent’s quality of character, *answerability*, which is conditioned by the agent’s quality of judgement, and finally *accountability*, which is conditioned by the agent’s quality of regard. Each of these types of responsibility also legitimate different syndromes of reactive attitudes toward the agent in question, namely: disdain/admiration, regret/pride, and anger/gratitude respectively. So, poor quality of character would make an agent an apt target for disdain, whereas a sufficiently good quality of regard would make an agent an apt target for gratitude, and so on. Importantly for my overarching argument, Shoemaker (2015a: 224-225) identifies accountability and the quality of regard as that type of responsibility and that facet of will where control plays a necessary role, though he understands this control differently to how the volitionists we’ve discussed above do – which is probably unsurprising given that Shoemaker’s overall strategy is still more akin to that of an attributivist. For him, the kind of control necessary for accountability is “*empathic control*”, and to have such control an agent must “take the facts about others’ normative perspectives as putative reasons or one’s responding in a simpatico emotional fashion to others is governed by one’s identification with them (i.e., these perceptions or responses are empathy-

sensitive).” By his own admission, Shoemaker does not expect this control condition to be considered sufficiently robust by those that argue that control is necessary for moral accountability, but hopes to deflate this worry by stressing that even the harshest reactive attitudes, anger and gratitude, do not necessarily imply “harsh treatment”, and so “the motivation to seek out a more robust conception of control will just diminish.”

Similarly, Watson argues that we should endorse pluralism to make sense of the ambivalence cases, but contra Shoemaker he argues for the recognition of *two* types of responsibility: *aretaic* and *accountability*. He describes how he sees the difference between these two variants as follows (2004: 266):

In one way, to blame (morally) is to attribute something to a (moral) fault in the agent; therefore, to call conduct shoddy *is* to blame the agent. But judgements of moral blameworthiness are also thought to involve the idea that agents deserve adverse treatment or “negative attitudes” in response to their faulty conduct. The former kinds of blaming and praising judgements are independent of what I am calling the practices of moral accountability. They invoke only attributability conditions, on which certain appraisals of the individual as an agent are grounded. Because many of these appraisals concern the agent’s excellences and faults – or virtues and vices – as manifested in thought and action, I shall say such judgements are made from the *aretaic perspective*.

For Watson then, the primary difference in which responses the two variants of responsibility legitimate is that only accountability seems to justify the usual reactive attitudes, whereas nothing further than the attribution of moral fault is justified by aretaic responsibility. And in terms of the conditions necessary for each variant, he argues that aretaic responsibility requires the conditions usual of a Reason-based real self view – in this case, that the behaviour be reflective of the agent’s valuational system – whereas accountability (as it involves the imposition of adverse treatment to its target) requires that the agent meet the “*control principle*” or some requirement of avoidability (2004: 274). This further requirement is roughly understood to be some version of *the power to have done otherwise*, and Watson is skeptical that it is ever met.

What can be immediately noticed is that both accounts have certain clear similarities, beyond the obvious fact that they are both variantist accounts. The notion of attributability that Shoemaker employs is very similar to Watson's aretaic responsibility, which is no accident given that Shoemaker discussed aretaic responsibility as an example of responsibility as attributability. They also both employ the notion of accountability, and both link this type of responsibility to control and adverse treatment. There are glaring differences as well: Shoemaker takes attributability to legitimate certain reactive attitudes, which Watson seemingly does not, Shoemaker takes quality of character to be the condition on attributability where Watson takes it to be the quality of the agent's valuational system, and Shoemaker introduces a whole new variant of responsibility in the form of answerability.

Despite these differences, both accounts do provide plausible answers to the problem posed by the ambivalence cases. It would be reasonable to think that they could dissolve any lingering confusion brought on by these cases: Skip and Harris are plausibly attributively responsible, but do not seem to be apt targets for accountability. And if employing Shoemaker's framework, they would also be answerable. Thus, these accounts allow us to capture and conserve our practices in these cases. However, this does come at the cost of rejecting invariantism. It is not clear to me that this is any great cost however, though as we will see, this may depend on how the relationship between the different variants are meant to be understood.

Confronted with the explanatory potential of variantist approaches to moral responsibility, a family of questions can be asked regarding how the relationship(s) between these variants are to be understood. Out of this family I take the two most significant questions to be: (1) what feature or set of features, presumably shared by each variant, allows us to consider the responsibility in question to be *moral* responsibility? And, (2) can the conditions for the different variants be jointly satisfied?

(1) is not infrequently raised as an objection to pluralist accounts. Levy (2005) argues that responsibility that is not of accountability variety – that is to say, responsibility that does not meet a necessary control condition – is too shallow or superficial to count as genuine moral responsibility. These criticisms mirror and expand on those

presented against real self views by Wolf, who also argued that such attributability approaches are too shallow to be accounts of moral responsibility. On the other hand, it seems open to attributivists – who want to argue that control does not play a necessary role in determining moral responsibility – to argue that accountability is not a unique type of moral responsibility. It can be thought that whereas attributability answers the question, “is this agent morally responsible?”, accountability is concerned with the secondary question of “what responses does this agent being morally responsible legitimate?” and so is not a type of moral responsibility. Either of these is a likely strategy for a proponent of invariantism to employ. To be clear, even if successful, these strategies would require the sacrifice of a degree of conservatism, as the ambivalence present in these cases would be attributed to an error among the folk. For Levy, we are simply wrong to think that Skip or Harris merit moral responsibility, whereas for an invariant attributivist both agents *are* simply morally responsible, with no explanation for why these cases seem to call for different responses than would be the case if Skip and Harris were normally functioning human adults with normal formative histories who performed these same behaviours. I will not present a defense here of either Shoemaker or Watson’s particular variants of moral responsibility against this objection,<sup>75</sup> but will discuss my defense of the variants present in my own pluralist account in the next section.

Turning now to question (2). There are at least three potential answers to this question, all of which have been defended in the literature, namely: one can adopt the position of certain hard determinists – who have always claimed that the conditions for moral responsibility are never satisfied – and argue that these new variantist approaches add nothing interesting to the discussion. Alternatively, one could answer that these conditions can all be satisfied independently (call this the Independent Satisfaction position), such that the presence of any given type of responsibility is wholly independent of the presence of any other. Shoemaker is an example of someone who endorses Independent Satisfaction. Finally, one could answer that the conditions of one type of responsibility is partially constitutive of the conditions of another, such that the presence of one type of responsibility is dependent on the presence of another (call this the Dependent Satisfaction position). Watson is a

---

<sup>75</sup> Both thinkers provide their own defenses against these objections, see Shoemaker, 2015a, and Watson, 2004.

proponent of this position. I will be ruling out the first answer as, though it may well be true, I will be developing my argument as though compatibilism were true, a la my strategy from Section 1.2.<sup>76</sup>

Between Independent and Dependent Satisfaction there is more of a trade-off. Independent Satisfaction helps to secure the intuition that the types of responsibility under discussion are indeed *discrete* (though related) variants of responsibility. On the other hand, if an account of this sort is understood as a brutally disjunctive one, it more immediately raises the concern that these variants do not share a sufficient foundation such that all can be considered as examples of *moral* responsibility – this is clearly related to question (1). Dependent Satisfaction, in contrast, can provide a clearer answer to the worry about why the different variants should be considered examples of moral responsibility, as these variants share at least some of the same conditions. Watson, for example, takes aretaic responsibility to be a condition for accountability: if a behaviour does not meet the attributability conditions required for aretaic responsibility, then that behaviour cannot be the basis for accountability. For Watson, the addition of a control or avoidability condition on accountability is a further condition. This approach also softens the objection that the account is disjunctive, as though one variant of responsibility, X, may be present without the other, Y, Y cannot be present without X. However, Dependent Satisfaction does give rise to the worry that the dependent variant – accountability in Watson’s case – does not represent a clearly distinct type of responsibility. This, of course, is a repeat of the invariant attributivist’s challenge from the discussion of (1).

Quite a bit of the plausibility of pluralism about moral responsibility depends on how one answers (1) and (2). As such, in the next section I lay out my own pluralist account, and then attempt to provide compelling answers to both questions.

#### *4.3. My dual-variant account of moral responsibility*

In this section I will be arguing that there are (at least) two distinct variants of moral responsibility, and that though these variants might share the same objects (agents), they have different conditions for occurrence, as well as legitimating different moral

---

<sup>76</sup> Such a hard determinist can read my arguments as speculation: “if compatibilism were true, then...”

responses. Although I will be employing a mix of Shoemaker and Watson's terminology in labelling these variants as *responsibility as attributability* and *responsibility as accountability*, I will be developing revised versions of both. It is my hope that such revision will capture the keen insights behind both thinkers' deployment of these notions, while improving upon them. As will be shown, the latter variant is inextricably tied to reason-responsive control whereas the former is not. It should be noted right off the bat that, given my overarching argument, I will not be giving a full account of attributability. Since my aim is to explain the relationship between intentional behaviour and moral responsibility, and I take intentional behaviour to require reason-responsive control, it is the variant of responsibility necessarily linked to control that I take to be key. However, to make the move to pluralism compelling, I will endeavor to say enough about attributability to make its distinction from, and relationship to, accountability palatable.

I begin by readily admitting that my account follows that of Watson more closely than that of Shoemaker. I take it to be an advantage for a pluralist account to only add additional variants when necessary to capture some fact of moral responsibility practice, and so if possible we should prefer a bipartite model over a tripartite one. If two can get the job done, then two are as many as are needed. Another advantage to employing two variants of responsibility is that, once these variants are properly fleshed out, it becomes possible to resolve the apparent tension between volitionism and attributivism (which was one of Watson's central motivations in originally arguing for the recognition of the two "faces" of responsibility). Simply put: each of these approaches is concerned with a different variant of moral responsibility. These are the reasons for my choice of a dual-variant account, to which I now turn.<sup>77</sup>

#### 4.3.1. *Attributability*

My use of the notion of responsibility as attributability adopts much from Shoemaker's use of the same notion, as well as Watson's notion of aretaic responsibility. This is the variant of responsibility involved in cases where I judge an

---

<sup>77</sup> This opens the way for the obvious objection that, using this reasoning, a unitary model would be even better. My response to this is that yes, such a model would be preferable if it were possible, but it is not. The entire point of the ambivalence cases is to reveal that no such unitary response can be adequate.

agent's moral conduct as faulty, and may have certain reactive attitudes toward him such as contempt or disdain (a la Shoemaker), and may further be inclined to alter my interpersonal relations as regards the agent. Scanlon (2015: 91-92) gives a good overview of the kinds of alteration to our interpersonal relationships this may include, such as: withdrawing my trust, decreasing my readiness to help him with his projects or emotionally invest in his success or failures<sup>78</sup> – all alterations that align with the attitude of disdain. Scanlon takes these responses to be reactive attitudes “in the general sense that Strawson defines: attitudes toward a person, including changes in one’s intentions about how one will treat or respond to him or her, that are adopted in response to that person’s attitudes toward oneself or others.” I am inclined to agree with him, but nothing in my account hangs on whether these responses are thought of as reactive attitudes. Agents can of course be praiseworthy as well as blameworthy on this account, and when this is the case the reactive attitude legitimated is that of admiration. Where the interpersonal relationships are concerned, we will see the opposite movements to what we saw with blame: an increase in trust, increase in readiness to help, and greater emotional investment in the agent’s successes and failures – again, changes that align strongly with admiration. Despite this symmetry, I will generally focus on cases of blameworthiness, as these have generally been the source of greater contention. Quite clearly these types of responses seem appropriate in cases such as that of Skip and Harris. When discussing a variant of moral responsibility, there are two important elements to consider: what are the conditions of this variant, and what responses does the presence of this variant legitimate? We have already been introduced to the responses legitimated by attributability, and so turn now to its conditions.

I take the condition for moral attributability to be some version of the conditions for responsibility put forward by QoW accounts. As such, rather than introduce my own condition of attributability, I will instead endorse MORAL EXPLANTORY BLAME (CREDIT) as the appropriate condition. To recall:

---

<sup>78</sup> Scanlon also includes in his list a decreased willingness to enter into special relationships such as friendship with the person. This may be a justified response, but I would argue that it is actually a result of the other three shifts in interpersonal relations. The reason that I would be less willing to be friends with such an agent is because of my withdrawal of trust, lack of willingness to support him, and lack of emotional investment in his success or failures.

MORAL EXPLANATORY BLAME (CREDIT): X deserves moral blame (credit) for Y if and only if Y is morally bad (good) and is explained in a normal way by X's quality of will falling below (above) what could be properly morally demanded of X

However, recall this condition only allows for differences in degrees of responsibility based on the moral valence of the given outcome and how much the degree of the agent's deviation from the normatively expected level of care is required in the explanation of the object of blame or praise. So, as it stands, this condition is not exclusively linked to those responses legitimated by accountability. In order to introduce this aspect into the condition, it is necessary to differentiate between those behaviours that are "explained in a normal way" involving control, and those without (I will be limiting my discussion to behaviour, given my overarching argument, but Björnsson certainly does not take his account to be limited to such, presumably attitudes and judgements could just as easily be accounted for). As we saw in Section 3.2., Björnsson argues that the perceived need for a control condition can be wholly accommodated by the requirement that the explanatory link between behaviour and quality of will must be a *normal* one. However, there is an ambiguity involved here: it can be asked whether or not the explanation of the behaviour was normal relative to the kinds of responses that we take to be legitimated. In other words, what would count as a normal explanatory link between behaviour and quality of will for legitimating the response of withdrawing trust, may be very different to what would count as such for legitimating the response of moral anger. Recall the stream of activity and the process of individuation. Just as there are many ways to individuate out activities relative to what is used as guidance for the individuation (such as intentional status for example), so too are there many ways to understand an explanation as normal relative to the purpose the explanation serves.

To see what the implications of this may be, consider this example from Björnsson (2017a: 147):

*Knockout:* Leaving the room, Victor pushes the door open quickly and with great force, inadvertently knocking unconscious the person just about to open the door from the busy corridor outside. At the moment of action, it didn't cross

Victor's mind that opening the door in that way might hurt someone, though he would have realized this if the question had come up

For Björnsson (*ibid.*: 158), Victor would be responsible if it were the case that his failure to appreciate the information available in memory and perception – “based on which Victor could have realized that he might be putting others in danger or distress” – was due to a deficit in concern. Björnsson makes the further point that Victor is responsible in this case despite not meeting the kind of control condition of the sort proposed by Levy (2017b: 152-153), and that it does not seem plausible to think that this could be a case resolved by tracing (2017a: 147). I agree that Victor is responsible in this case, and that this is despite Victor not having conscious control over the outcome, but my contention is that this is a case of *attributability* only, and Victor's behaviour legitimates only the responses that follow from this. In order for it to be the case that Victor be open to those responses legitimated by accountability – which I will discuss in the next section – it is in fact necessary that he meet some control condition (though not necessarily conscious control). It is on this point that I hope to expand on the explanatory quality of will account by illuminating this second, more stringent understanding of what it means for an agent's behaviour to be explained in a normal way by the agent's quality of will.

#### 4.3.2. *Accountability*

Holding an agent accountable involves strong reactive attitudes such as anger and indignation, and legitimates, merely on the basis of the morally relevant behaviour, sanction in the case of moral blameworthiness. These responses then involve the adverse treatment of the agent being held accountable – not merely in that it alters our personal relationship to them, as attributability legitimates, but that, as Scanlon (2015: 90) might say, it changes our *obligations* toward them. I may no longer be obligated to save an accountable individual from certain harms, or may be justified in taking certain actions that in fact impair the agent's liberty without violating obligations that I would otherwise have not to do so. The presence of such responsibility could, for example, justify the purposeful removal of liberty from some blameworthy individual, purely as desert for his or her behaviour. It is the fact that this variant of responsibility legitimates these responses that places a demand on an account thereof to meet the

concern of fairness, which – as Watson identified – entails the need to include a control condition. On the positive side, if an agent is praiseworthy in the accountability sense then this legitimates the reactive attitude of gratitude. It will also legitimate changes in my obligations toward the agent, but in this case, it might mean that I come under new obligations that hadn't been there before: e.g. if someone goes out of their way to return my missing wallet to me rather than keep it for himself or simply ignoring it, then it seems reasonable that if I notice that he has dropped his wallet I am more obliged to return the favour than would a bystander. It is often the case that meeting these obligations strongly aligns with our gratitude, and this does not seem to be an accident, any more than the fact that the presence of moral anger aligns with the negative alteration of obligations in the case of blameworthiness. In both directions, the reactive attitude and the alterations to obligation legitimated are complementary to each other, though the former is a conative state and the second concerns what we take to be normative reasons for action.

As such, the condition on accountability is stricter than that on attributability. The understanding of “explained in a normal way” must in this case be:

*Strict Interpretation:* for some behaviour, X, to be explained in a normal way by the quality of will of an agent, Y, such as to justify the moral responses characteristic of accountability, the behaviour must be under *normal System 2 Oversight*.

What this interpretation clearly does is to introduce a control condition, one that is related to the control condition on intentional status that was developed in Chapter 2. In line with the discussion of different volitionist positions in Section 3.1., the notion of control here involves reason-responsiveness, but *not* conscious control. It also incorporates an epistemic element since System 2 Oversight has inbuilt epistemic requirements. What is different about this notion of control as compared to that captured by S2O, however, is the addition of the qualifier, “normal.” I will explore this in the next section, where I develop in detail how this control condition is to be understood.

Having outlined my two variants of responsibility, I now tackle answering questions (1) and (2). Beginning with (1): what binds both variants together, what makes them both variants of moral responsibility, is, as Scanlon says (2015: 108) in discussing his own two variants of responsibility:<sup>79</sup> “They are both properly called forms of responsibility because they both assign moral significance to what an individual is like or has done.” Another way to grasp this is to consider the similarity that was identified between volitionist and attributivist accounts in Section 3.3., namely: that both approaches take moral responsibility for behaviour to fundamentally be a matter of responding to some revealed morally relevant feature of the agent. In the case of both variants, the assignment of moral significance depends upon the way in which the agent’s behaviour is explained by the agent’s quality of will, where they differ is in what must be involved in this explanation. It is also the case that both variants legitimate commonly recognized moral responses, which is most pertinent for attributability, as it reinforces it against the criticism that it is not a genuine kind of moral responsibility. In response to this criticism from the volitionist direction, that attributability is not properly moral responsibility, I am inclined to agree with Björnsson’s contention that such positions are so at odds with our responsible practices, and would require such a sacrifice of conservatism, that this should be counted as a disadvantage to their approaches.

Question (2) asks after the relationship between the variants in an account, and in the case of mine this relationship is one of Dependent Satisfaction. Accountability can only be present if attributability is present. This follows from the fact that the condition on accountability is simply stricter than that on attributability by adding the requirement for control. So, accountability will be present in any case of attributability where the control condition is met. This raises the worry that it might not be clear why accountability should be seen as a distinct type of responsibility. However, this does not seem a strong objection, given that the different variants legitimate different – and widely recognized – *moral responses*, and do have different, even if partially shared, conditions. It is also the case that whereas the responses legitimated by attributability are somewhat passive, those legitimated by accountability “implicate *confrontation* of a kind” (Shoemaker, 2015a: 87). This

---

<sup>79</sup> *Moral reaction and substantive* responsibility (Scanlon, 2015).

confrontation should be understood as communicative, it aims to communicate moral anger and indignation to the target, without necessarily implicating harm to him or her. This kind of call to confrontation is not present in cases where only attributability is present.

As I have mentioned, I take it to be the case that attributivist accounts of moral responsibility have been largely aimed at explaining the conditions for moral attributability, whereas volitionists have been concerned with the conditions for moral accountability. It should therefore be no surprise that, as long as the assumption of invariantism is retained, these two approaches would find themselves in tension over the role that control should play in the condition on moral responsibility. By giving up on this assumption, it becomes possible to embrace the keen insights presented by both approaches. I take this possibility to count in favour of pluralist accounts of the kind I have sketched. Even so, this is far from a complete resolution of this tension – and such an enterprise exceeds the scope of my arguments. It is unlikely, even if invariantism is rejected, that volitionists and attributivists will agree to the formulation of the two variants I have put forward, or perhaps with how I understand the relationships between them. After all, within each approach there are important questions: “what kind of control is necessary for responsibility?” and “which feature of an agent is the morally relevant feature that must be disclosed?” for example. But at the least it presents the hope of a resolution.

Another advantage of endorsing this pluralist picture is that it helps us to make sense of the case of the panicked skipper who jettisoned his cargo raised by Fischer and Ravizza in Chapter 2: Section 1.1. In the discussion of that case, I compared a skipper who was truly driven by an irresistible fear to act as he did to a person emerging from a truly blind rage for the first time, and claimed that in both cases the actions undertaken should not be counted as intentional, as they were not under appropriate control. It should then be no surprise that I do not take these agents to be fitting candidates for accountability. They may well be open to attributable responsibility, however, if it were the case that their behaviour is explained in a normal way (understood in the looser sense, where control is not relevant) by their poor quality of care. This will depend how the details of the case are specified, as the degree to which

each agent's situation reduced their capacity to care will determine both whether the agent is attributably responsible, and to what degree.

#### *4.4. Concluding remarks*

In this section I have introduced the notions of invariantism and conservatism, and shown that they are in tension. I proceeded to argue that given the existence of so-called ambivalence response cases – those where certain types of moral response feel justified while others certainly do not – no invariantist conception of moral responsibility can adequately maintain conservatism. In light of this, I followed Watson and Shoemaker in arguing in favour of a pluralist approach to moral responsibility. After investigating the pluralist accounts of these two thinkers, I identified two important questions that such accounts must answer. The reason being that the nature of these answers does much to determine the persuasiveness of the given account. I then presented my own dual-variant account of moral responsibility, which takes there to be two variants: attributability and accountability. The crucial distinguishing feature between these two variants is that the latter has a control condition, whereas the former does not. In laying out my understanding of these conditions I borrowed heavily from the explanatory quality of will account of moral responsibility. I then attempted to improve upon it by introducing a pluralist element in the form of two interpretations of what it means for a “behaviour to be explained by an agent's quality of will in a normal way”, and attempted to answer the two questions identified prior.

In the next section I continue the development of my account by fleshing out what I take to be the control condition on moral accountability: *normal System 2 Oversight*. And once this is done, show how the intentional status of an agent's behaviour and an agent's moral accountability for that behaviour are necessarily linked through what I call the *nexus of control*.

## 5. The nexus of control

### 5.1. Control condition on accountability: *normal System 2 Oversight*

As my discussion of reason-responsiveness in Chapters 1 and 2 revealed, there are several distinct similarities between the control necessary for the exercise of intentional agency and the control necessary for moral accountability. In both cases the presence of this control is necessary (and in the case of intentional status, I take it to be sufficient) for a given behaviour to be reflective of an agent's autonomy. The kind of control in question must be both reason-receptive and reason-reactive (though the degree required may differ). And lastly, there is an epistemic component to this control, but it falls short of a demand for conscious awareness. It is therefore unsurprising that those who endorse the role of a control condition on moral responsibility (i.e. the volitionists) would be drawn to limiting the scope of responsibility to that which is done intentionally. However, adopting this position is untenable, given the cost to conservatism. But once we adopt a pluralist approach, it can be readily seen that accountability, which does demand control, may well be limited in this way. This indicates that System 2 Oversight will again be a useful way to understand the kind of control at stake.

This is not to say that the set of behaviours that are intentional is coextensive with the set of behaviours for which an agent is accountable. As has been previously noted there are many things that we do intentionally for which we cannot be held accountable. This is because, despite the similarities, there are also important differences between the nature of control required for intentional status and that required by accountability. I argue that this difference can be captured by the difference between S2O and *normal* S2O, and by the fact that the presence of S2O is both a *necessary and sufficient* condition for positive intentional status, normal S2O is by itself *only a necessary* condition on moral accountability. In the next section, I turn to unpacking what exactly is meant by this notion of normalcy. But first I must consider two of Björnsson's arguments against the inclination to understand the requirement that behaviour must be explained in a normal way in terms of control. The first argument, which is very much aimed at those who require the presence of conscious control, is that it seems that we frequently hold agents responsible for non-

akratic behaviours, which such a stringent demand would seem to rule out. The second involves the case of *The Catch*, and is a more general objection applied to accounts that take the presence of some appropriate notion of control to be necessary and sufficient for moral responsibility.

My response to the first objection is to agree that the demand for conscious control is simply too stringent, and would require a very significant revision of our moral responsibility practices. The kind of control that I seek to introduce is not control of this variety, but rather control through the *oversight* of the agent's dual process mechanism. Further, the underlying assumption of this objection, that conservatism should be taken seriously as a virtue for an account of moral responsibility, I take to be a motivation in support of my own pluralist position, as without the recognition of the need for a variantist understanding of moral responsibility we would not be able to capture the fact that our responsibility practices call for different responses in different cases, not only in degree but also in type, and that it seems that these different types demand different conditions.

The second objection is more troublesome. Björnsson introduces an example that he takes to show that even when control is present – and in this case full-blown conscious control – this is still not sufficient for moral responsibility. Rather what is necessary is the explanatory role of the agent's quality of will (2017b: 156):

*The Catch:* There was only one way in which you could prevent the deadly explosion, and you knew it. If you were to press the button, you would prevent the explosion, *but only if you would not be acting for ultimately moral reasons, out of concern or respect for potential victims.* (Those meddling neuroscientists were at it again, monitoring your deliberation and tracking your motivation.) Moreover, you were able to press the button, and able to prevent the deaths by doing so for non-moral or immoral reasons: perhaps to listen to the somewhat interesting sound emitted by the button, or to save the explosive device for an even deadlier occasion later on. As it happened, however, you did not at the moment care about those other things. So you did not press the button.

My response is to claim that this would only be an objection to my account if my argument was that the presence of control is *sufficient* for moral responsibility. This, however, is not my claim. Instead I argue that the presence of control is merely *necessary* for the presence of accountability. What I would argue is that the presence of a deficit or surfeit of good will is also in the same way necessary, though not sufficient, for moral accountability. Both elements are necessary and jointly sufficient in order for an agent to be morally accountable for a behaviour. What is certainly true is that when the correct explanatory link exists between the quality of will and the behaviour, this is sufficient for moral responsibility. However, in the particular case of moral accountability, this link must necessarily involve control.

### 5.1.1. *Normal functioning*

There are undeniably times when an agent acts intentionally but is an unfitting candidate for being *held* accountable. However, such cases come in three different forms. In the first the agent *is open* to moral accountability on the basis of some behaviour, but is not praiseworthy or blameworthy because this behaviour is morally neutral given the moral reasons in the vicinity. For example:

*Reader:* Megan, a young doctor, comes home from a day's work at the hospital and after getting comfortable sits down to read the novel *Ender's Game* in her spare time.

Megan is neither praiseworthy nor blameworthy for her reading of the novel, if there wasn't a sufficiently weighty moral reason for her to stop doing so – such as a need to call and check up with her ill mother for example.

The second variety is where the agent's behaviour is morally valent, and intentional, but the agent lacks sufficient control for moral accountability, and so *is not open* to it:

*Psychopath:* Skip – from the ambivalence cases – breaks a woman's arm when she resisted his attempt to force her onto his lap. His response is to blame the woman for resisting, and he feels no remorse for his actions.

Skip may well be attributably – and, for the record, legally – responsible for breaking the woman’s arm, but he lacks a sufficient capacity to be responsive to reasons to be open to moral accountability. More particularly, he lacks a sufficient capacity to respond to moral reasons, as a result of his insufficient capacity to care, which in turn is the result of his lack of empathy. This relationship between the capacity to care and the capacity to respond to moral reasons is an important one. My talk of control has largely been couched in terms of responding to reasons, and it is just such a notion of control that I seek to plug into MORAL EXPLANATORY BLAME (CREDIT) in order to provide the condition on moral accountability. Yet, the explanatory quality of will account is couched in terms of *caring* or *moral concern*. How these two ways of talking about the subject matter relate needs to be clarified.

My view is that the capacity to care<sup>80</sup> is necessary for the capacity to respond to moral reasons (as moral reasons). Importantly, this does not mean that a failure of an agent to *actually* care in a given situation undermines that agent’s capacity to respond to reasons. Rather, changes in an agent’s *actual* degree of care will influence the agent’s *actual* responsiveness to moral reasons. It is important to separate these out: an agent can have the capacity to respond to a certain moral reason, but fall short of this standard. On my view, openness to moral accountability is only jeopardised when the *capacity* to respond appropriately to moral reasons falls below a certain threshold. A reduction in actual empathy that reduces the actual responsiveness that is not accompanied by a reduction in the capacity for care, and so a reduction in the capacity for reasons-responsiveness to moral reasons, does not prevent openness to moral accountability. It is also the case that reductions in capacity to respond to moral reasons may reduce the degree of responsibility if this reduction explains the agent’s actual responses. This still leaves the threshold below which the capacity to respond to moral reasons must fall for an agent to no longer be open to moral accountability undefined, and I will rectify this soon, but first let us look at the third variety of intentional behaviour where moral accountability is absent, as it relates directly to foregoing discussion.

---

<sup>80</sup> We can substitute “care” for “moral concern” or “an intrinsic desire for the right or good” here, and I take the point to be unchanged.

These final cases are ones where the agent has no capacity to respond to moral reasons. For example:

*Ideal psychopath*: Judith is a psychopath. What is more, she is an *idealised* psychopath – as was also introduced in Chapter 2: Section 1.1. – and so, is *incapable* of responding to moral reasons at all. Judith is bored one Friday night and because it gives her a thrill, traps the neighbour's dog and kills him.

There may well not be any actual adult humans that fit this description, but it remains an important theoretical possibility. Importantly, Judith is not the same as a non-human animal since she can still respond to an otherwise normal spectrum of non-moral reasons. She is also different from a child, who may lack the capacity to respond to moral reasons, but who possesses (in the clear majority of cases) the chance to develop such a capacity. Due to her incapacity, and presuming that it's permanent, Judith will never be open to moral accountability. Indeed, I am inclined to concur with Fischer and Ravizza (1998: 82) that an agent such as Judith does not qualify as a moral agent.

Given the discussion of these cases, it seems clear that the capacity to respond to moral reasons beyond some threshold is a necessary condition for moral accountability. I will call this capacity *Moral Competency*. But how does the demand for moral competency relate to the control condition on moral accountability? Very naturally, or so I will argue. A human agent's System 2 oversight mechanism, if it is functioning normally, will have the capacity to be responsive to moral reasons beyond the mentioned threshold. The reason we should think this is the case can be illustrated by the very fact that agents who are sufficiently unresponsive to moral reasons – such as Skip and Judith – are both agents where we would say that they have *abnormal* reason-responsive mechanisms. But to make this point more clearly, we must unpack the notion of normalcy.

The use of normalcy here is technical, not colloquial. For a process or mechanism to function normally is for it to function sufficiently in line with its operative characteristics, where these operative characteristics are understood as normative expectations constitutive of the mechanism, and “sufficiently in line” should be

understood as a bandwidth. A dual process mechanism *is* a reason-responsive mechanism, that it be reason-responsive is a normative expectation we have of it. For such a mechanism to function normally then, means that it must have the capacity to be receptive and reactive to reasons within a certain bandwidth (let's call this the *Normal Band*). The functioning of two mechanisms – or one mechanism at different times – may deviate from each other within the *Normal Band*, yet still be normal provided they still fall within it. In this sense, normalcy comes by degrees.

Another aspect of normalcy is that raised by Martin Smith (2010: 15), that whereas abnormal states of affairs call for *special explanations*, normal states of affairs do not. This is abundantly clear in the cases of Skip and Judith, where the functioning of these agents' System 2 oversight mechanism calls for a special explanation, precisely because of their lack of capacity to respond to moral reasons. However, note that using this notion of normalcy an agent can have a normal capacity to respond, yet fail to actually respond consistently. To see this, consider the case Smith uses of the man driving home from work:

Other times—when we say things like ‘Tim would normally be home by six’ or ‘When I turn my key in the ignition, the car normally starts’—part of what we are trying to express, I believe, is that there would have to be some satisfactory explanation if Tim wasn’t home by six or the car wasn’t starting.

In this sense of “normal” it could be true that Tim is normally home by six, even if this occurrence is not particularly frequent. What is required is that exceptions to this generalisation are always explicable as exceptions by the citation of independent, interfering factors—his car broke down, he had a late meeting etc. If this condition is met, then the best way to explain Tim’s arrival time each day is to assign his arrival by six a privileged or default status and to contrastively explain other arrival times in relation to this default. (ibid.: 15-16).

It is also important to note that the operative characteristics, though being normative expectations, are not *ideal* expectations. A coffee machine does not always function more normally the closer it comes to the perfect execution of its operative characteristics. Similarly, a certain dish of food might have, as an operative

characteristic, a certain level of flavour. This level is not “ideal” flavour, or perfect flavour, or “the most flavour that the dish could have”, but rather the level normatively expected of the dish. In the case of the dish of food, certain instances may be more flavourful or less flavourful than this expected level, and if they were sufficiently so then special explanation would be called for: “Why is this soup so terrible!” or “What did you do with this, it tastes great!”. At the same time, dishes within the *Normal Band* can still be discriminated between in terms of their distance from the expected level: “That was a pleasant soup” or “That was a less flavourful than I expected”. The dishes in these cases are both still normal, but they are less normal than an instance that falls closer to the expected level: “this is how it normally tastes”. What is true of dishes of food holds also for coffee machines or dual process mechanisms: the closer that the reason-responsiveness is to the expected levels, the more normal its functioning, and if it falls sufficiently away from this level, then the functioning is abnormal (super- or subnormal).

Now, in the case of the functioning of the System 2 oversight mechanism, there are two dimensions of normalcy that could be violated: the first I will call *Normality in Exercise (NiE)*, and the second *Normality in Capacity (NiC)*. Consider:

*NiE*: an agent could simply be failing to respond to moral reasons that we would normally expect an agent in her context and with a normally functioning System 2 to respond to. She is abnormal in her exercise of her capacity to respond to moral reasons. I take violations of this sort of normative expectations to constitute moral failings, and that our demand for agents to answer for them goes hand-in-hand with the fact that they call for special explanation. I take it that in most cases such explanations will bottom out in a lack of (though not a lack of capacity to) care, usually because of the influence of the agent’s egoistic reasons.

*NiC*: an agent could be failing to respond to moral reasons that we would normally expect an agent in her context to respond to, and the explanation for this is that the functioning of her System 2 violates our expectations in that either: (i) it lacks the capacity for System 2 oversight, or (ii) the reasons-responsiveness to moral reasons falls below the *Normal Band*. She is abnormal in her capacity to respond to moral reasons. I take violations of this sort to undermine openness to moral accountability. I

take it that explanations for such violations will usually bottom out in stories about the agent's psychological history and capacities (e.g. an abusive childhood, a significant mental handicap, etc.). When I speak in this dissertation of a normally functioning System 2 oversight mechanism, I mean in this dimension. Also note that for an agent to be open to moral accountability, they need have *at least* a normal capacity to respond to moral reasons. Agents with a supernormal capacity remain open to accountability.

For us to know whether an agent violates NiE or NiC in a given case, particularly when we are looking at that case in isolation, will usually (if not always) require further examination. However, I take it that in the vast, vast, majority of cases the violation will be of NiE, and that this is the reason why we usually assume someone to be exhibiting a moral failing, rather than a lack of moral agency, in almost all cases. This helps to highlight that the practice of holding an agent responsible is a *process*. As Rosen (2003: 61) says,

morality, like the law, operates with a defeasible presumption of responsibility. Suppose Jane does something wrong. Suppose she steals a candy bar from the corner store. Until we hear more we are entitled to suppose that she is liable to blame for what she's done. But if it turns out that she is only five years old, or that she was coerced, or that she has just contracted kleptomania, we may conclude that even though the act was wrong, it would be a mistake for us to blame her or for her to blame herself.

Where Rosen says that we are entitled to suppose that Jane is open to blame, I would say that it is *understandable* for us to provisionally judge her to be to blame, and to even respond as though she is. However, we are morally obligated to discover whether our provisional judgement is in fact correct, and so whether Jane is actually a fitting candidate for blame. I take McKenna (2012) to be providing an account of this process in his *Conversational theory* of moral responsibility.<sup>81</sup> In this theory, he argues that there are three usual steps in the process: firstly, when an agent acts, and her acts are morally charged, she appreciates that she might be introducing a

---

<sup>81</sup>As a volitionist, when McKenna talks about responsibility, I take him to be talking about accountability particularly.

meaningful contribution to a kind of conversational exchange with others. This initial stage McKenna calls *Moral Contribution*. The second stage, in which that agent is blamed (or praised) by a respondent on the basis of the morally charged action, he calls *Moral Address*. In the third stage, *Moral Account*, the blamed agent extends the conversation by offering an excuse, a justification, or an apology, for example. The respondent might at this point continue the conversation, perhaps by forgiving or punishing.<sup>82</sup>

My own understanding of the process is similar to McKenna's though not identical. On my view, Moral Address involves a *provisional*, and defeasible, judgement and response on the part of the respondent. And Moral Account is not only a matter of the agent putting forward excuses or justifications, but also of the respondent investigating the facts of the situation – most relevantly here the normalcy of the agent's reason-responsive mechanism. And I take there to be a further step, call it *Final Address*, which is when the respondent arrives at their final judgement and response to the agent. Note that this process can break down at any step, and that the final judgement and response are by no means infallibly correct. There is, however, a pro tanto moral reason for the respondent to arrive at an accurate judgement – and so pursue Moral Account. A failure to do so can be morally blameworthy itself. At the same time, there is a pro tanto moral reason for the agent to meet the respondent's address and account for her or himself when so confronted. Of course, the moral reasons operative in these cases may be outweighed by other reasons, but they are there. This is a helpful way to understand what's going on in what could otherwise be confusing cases: the woman whose arm Skip breaks would, for example, be entirely understandable in holding him accountable during the step Moral Address. However, in the course Moral Account she should, and is likely to, recognize that Skip's reason-responsive mechanism violates NiC, and this should lead her to form the final

---

<sup>82</sup> This is somewhat akin to Gormally's (2016: 281-282) description of responsibility as a three-level concept, by which he means:

It may be used in reference, firstly, to being a cause or contributory condition of x happening; secondly, to being callable to account for x happening; and, thirdly, to being guilty for the occurrence of x, i.e. when one lacks an exonerating answer when called to account for causing or contributing to x happening.

Though Gormally's description of the first of these levels differs from McKenna's notion of Moral Contribution in that he seems to be identifying this level with *causal* responsibility, the overall structure of the three levels is strikingly similar to McKenna's Conversational theory, with the second and the third levels having clear commonalities with Moral Address and Moral Account.

judgement that Skip is attributably responsible exclusively, and respond appropriately.

Having laid all this out, I can now present my formulation of the control condition on moral accountability:

**Normal System 2 Oversight (nS2O):** an agent's X'ing, undertaken at time t, is under normal S2O, and so a legitimate basis for moral accountability, iff:

- (1) At time t the agent X'd
- (2) X is under some degree of guidance by the operation of (a) System 2 or (b) System 1 with System 2 oversight, such that the agent could bring it about with sufficient reliability through this guidance at time t
- (3) *The System 2 oversight mechanism is functioning at least normally*
- (4) The agent has a belief that she will at least try to X

A key feature of this control condition is that it does not posit any ownership condition in the vein of Fischer and Ravizza. The reason for this is that I share Björnsson's view that – when dealing with normative responsibility – why an agent has a given capacity is irrelevant, what matters is that the capacity is present, and the degree of its presence. This allows for responses to manipulation cases, as was discussed in Section 3.2. This condition also clearly rules out both Skip and Harris from accountability, as in both cases the agents presumably do not meet (3).

However, for the reasons discussed in Chapter 2: Section 1.5., this condition only applies to activities and effects, not to consequences. In the next section I argue that, just as with intentional status, openness to moral accountability can be transferred to the consequences of activities.

#### *5.1.2. Accounting for consequences*

To recall, I have argued that intentional status transfers from intentional activities to intentional consequences via:

**Transfer of intentional status (TIS):** a given consequence, X, of an agent's activity, Y, occurring at time t, is intentional iff:

- (1) The agent has the belief that bringing about X is, given her evidence, a possible part of the fulfilment of an aim to Z, which involves Y'ing
- (2) X is brought about under sufficient guidance of System 2 Oversight by Y'ing, in line with the agent's plan to Z

As may be guessed, I argue that openness to accountability can be similarly transferred via:

**Transfer of openness to accountability (TOA):** an agent is open to accountability for a given consequence, X, of the agent's activity, Y, occurring at time t, iff:

- (1) The agent has the belief that bringing about X is, given her evidence, a possible part of the fulfilment of an aim to Z, which involves Y'ing
- (2) X is brought about under sufficient guidance of *normal System 2 Oversight* by Y'ing, in line with the agent's plan to Z

Probably the most significant ramification of TOA is that moral accountability cannot extend to foreseeable but unforeseen consequences. Such consequences can also not be intentional, and the clearest way to understand why is to consider again the example of Victor from *Knockout*. In this example, having knocked down the person on the other side of the door, Victor is suddenly being blamed by her. In response Victor declares, “I didn’t do it intentionally!” And this seems to be a plausible response, we do not think of such consequences as intentional. On the responsibility side, though his declaration may not exonerate Victor entirely, it will at the least shift the type of responsibility from accountability to attributability. However, if it was the case that Victor has made a habit of such reckless door openings, and become aware of this fact yet has taken no steps to remedy this habit, then he will be open to moral accountability *for the failure to take remedial steps*. This will be a case of accountability for an omission – a case of negligence.

This would also be an example of tracing. Though my account does not employ tracing as a means of accounting for cases of responsibility where control is absent – I take my endorsement of a dual-variant account to see to that – it remains the case that in at least some situations an agent might bring about a consequence that is unintentional and not itself open to accountability, but where the agent might still be open to accountability for some preceding activity that was intentional.

### *5.2. Concluding remarks*

This then is the answer to the question of how moral responsibility relates to the intentional status of activities and outcomes: *an agent can be morally accountable only for those activities and outcomes that are intentional*. This follows from the shared role played by a kind of reason-responsive control best understood as System 2 Oversight. However, an agent is not open to moral accountability for all her intentional activities and outcomes, which follows from the fact that moral accountability has a stricter control condition than intentional status, namely: *normal* System 2 Oversight. What an agent is accountable for is also not exhaustive of what she could be morally responsible for, since my account endorses a dual-variant model of moral responsibility. So, behaviours that may be closed to accountability – such as unintentional activities and outcomes – may still be open to responsibility understood as attributability.

This unknotting expedition was originally motivated by the consideration of Anscombe's concerns regarding Truman's responsibility for the innocent deaths that resulted from the deployment of the nuclear bombs, and the viability of the response that since these deaths were unintended he should be exculpated (or have greatly reduced responsibility). Now that I have developed a solid account of the relationship between moral responsibility and intentional activity (and the outcomes of such), in the next and final section, I will attempt to see what consequence it may have for understanding this motivating example.

## 6. Consequences for the Doctrine of Double Effect

Anscombe's protest was based on the claim that honouring Truman was comparable to honouring Genghis Khan, Nero or Hitler, as he was a war criminal for the use of the atomic bombs. More particularly, for bringing about the deaths of many innocents (including children literally boiled alive in bath tubs) by dropping the weapons on Hiroshima and Nagasaki. Anscombe would, presumably, have not had these same reservations had the bombs been deployed on purely military targets, such as an enemy fleet. One line of opposition to her view about Truman's moral responsibility for the bombings was that the former president was not morally responsible for the bombings, or at least not in the way or to the degree that Genghis, Nero, and Hitler were for their crimes, because he did not *intend* to bring about the death of the innocent victims in the two cities. This defence is not a novel one, and is an example of the general strategy of the Doctrine of Double Effect (DDE), which roughly holds that "other things being equal, harm that is strictly intended is harder to justify than harm that is merely foreseen" (Woppard, 2017: 142). On initial brush, it may look as though my dual-variant approach would agree with the DDE defence, if it is read as claiming that: an agent is accountable for what is intended, but can merely be attributably responsible for what is merely foreseen. However, this answer is too swift. On my account, it is not only what is *intended* (strictly or otherwise) that can be open to accountability, but also all behaviour that is intentional. And such behaviour includes many foreseen (but unintended) effects and consequences of intentional activities. To get to an answer then we will have to dig deeper, starting with a fuller understanding of the DDE.<sup>83</sup>

There are many different formulations of the DDE.<sup>84</sup> Given this background, it is not my intention to propose a new formulation of the DDE, or to argue for new constraints or conditions to be added to it. However, if I am to discuss the subject I will have to adopt some formulation or other. In trying to reconcile these two realities, I have chosen to simply adopt Delaney's formulations of the DDE, as he is a strong

---

<sup>83</sup> Note, it is not my intention to deliver any final verdict on Truman's fateful decision. Even if the DDE fails to render his decision permissible, there may well be other considerations that do – such as lesser evil justifications. My interest is in whether the outcome of his decision, i.e. the bombings themselves, being intended or merely intentional is morally relevant.

<sup>84</sup> Some good examples are those of Mangan, 1949, McIntyre, 2001, and Masek, 2010, though there are many others.

defender of the DDE, and is not one of those thinkers who wish to either constrain the Doctrine<sup>85</sup> or revise it away from its roots.<sup>86</sup> It is my hope that by doing so the thrust of my argument can hit home against the strongest opposition.

The DDE is classically formulated as a set of conditions that help in adjudicating the moral permissibility of an action. In most such formulations of the Doctrine there are four stipulated conditions, which if met means that an action is morally permissible, or at least more permissible than an otherwise identical action that fails to meet these conditions. These are: (1) the act is not bad in itself, (2) the act also issues in proportionally good results, (3) that the bad result is not intended but merely foreseen, and (4) the bad result is not a means to the good result (Mangan, 1949; Quinn, 1989). All the conditions must be met in order for an action to be permitted by the DDE.

Delaney (2008: 335-336) offers what he calls a more “modern” formulation that runs as follows:

[I]t is sometimes morally worse to act with the intention to produce a bad effect as a means to a good end than to act while merely foreseeing that an equally bad effect will come about as a byproduct of one’s endeavouring.

In his notes (*Ibid.*: 359-360) he also provides the following, more precise, formulation:

[A]n action may be morally permissible if (1) the end is good (2) the means is at least neutral (3) the foreseen bad effect is not directly intended and (4) the foreseen bad effect is proportional to the good end after which the actor strives.

This formulation of the DDE captures what Masek (2010: 567) considered to be the most pertinent element of the DDE, the claim that “someone who causes bad effects must have both a good intention and a morally acceptable reason for permitting the bad effects.”

---

<sup>85</sup> See Anscombe, 1982 and McIntyre, 2001. Both these thinkers, in different ways, propose understandings of the DDE which would limit the range of cases to which it is applicable.

<sup>86</sup> I include Woppard here, as though she is a strong advocate for the DDE or some replacement thereof, for her “a replacement for the DDE need not appeal to the agent’s intentions” (2017: 143). As my interest is directly on the role played by intention, this is the variety of the DDE that concerns me.

For the purposes of my investigation, the focus is on (3) in Delaney's account, which posits a crucial morally relevant difference between acting "with the intention to produce a bad effect" and acting "while merely foreseeing that an equally bad effect will come about." Likewise, for Masek the requirement is that the agent must have a "good intention." In his discussion of what he means by "good intention," Masek clarifies that, for her intention to be good, the agent must not *intend* the bad effects of her actions. This becomes particularly clear when considering that he argues that a key challenge for a defender of the DDE is to "distinguish intended effects from side effects (or unintended effects)" (Masek, 2010: 567). In other words, for both Delaney and Masek, the distinction between the intended effects and the merely foreseen side effects<sup>87</sup> of an activity – and the difference in openness to moral responsibility that they take to track this distinction – continues to play the key role in the plausibility of the DDE. For shorthand, I will be referring to this set of ideas as *Distinction* from here on out.

My argument is that *Distinction* cannot be maintained. The distinction between outcomes that are intended and those that are merely foreseen does not necessarily, nor even characteristically, track a distinction in moral responsibility, of either type or degree. This is not to say that intended outcomes and merely foreseen ones are *always* alike in openness to moral responsibility, even if they are the same outcome. There is a morally relevant distinction in the vicinity (a few in fact, but one of particular note):<sup>88</sup> the distinction between outcomes that are *intentional*, and outcomes that are merely foreseen. Provided that the facts of the case are otherwise identical, an intentional outcome, X, is equally open to moral accountability whether it is intended or intentional – and likewise for attributability. Recall that the difference between an intended outcome and a *merely* intentional outcome (following from Bratman, in Chapter 1: Section 3.1.) is that the former requires an *intention* to bring about that outcome, whereas the latter does not *need* an intention to be present.

---

<sup>87</sup> My use of the term side effects will follow that of Masek, so far as it applies to the unintended effects of actions. I am not entirely certain that this is the most adequate use of the term, but for the sake of clarity I will employ it in this way for the duration of my arguments presented here.

<sup>88</sup> One such distinction is that between *doing* and *allowing*, often called the Doctrine of Doing and Allowing (DDA). As the role of intentional status is not used as a discriminator in this doctrine I will not consider it here, but Woollard provides an excellent overview and defense of the DDA in Woollard (2012a; 2012b; 2017).

Since moral attributability does not involve any necessary role for intentional status, I take it for granted that the DDE will have no bearing on this variant of responsibility. Looking at accountability, however, where positive intentional status is a condition for a behaviour to be open to responsibility there is reason to think that the DDE may well play a role in determining responsibility. So, if we assume that what is morally relevant for the purposes of moral accountability is whether the behaviour in question is explained in the normal way by the agent's quality of care – and this “explained in the normal way” is understood under the strict interpretation – then to think that what is intended and what is merely intentional marks a morally relevant distinction implies that an agent's intended outcomes are explained by the agent's quality of care to a greater degree than her merely intentional ones. I argue that this is not necessarily the case. A merely intentional outcome may be explained by an agent's quality of care to at least the same degree as if it was intended:

*Intentional*: Paul is caught in the classic trolley problem. Paul, as it happens, is a committed Utilitarian, and thinks that, as a general rule, it is always best to try to save the greatest number of people. He forms the intention to pull the lever and does so, regretting that his action will bring about the death of the single man tied to the track, but thinking it for the best.

*Intended*: Pauline is caught in the Fatman version of the trolley problem. Like Paul, Pauline is a committed Utilitarian, and thinks that, as a general rule, it is always best to try to save the greatest number of people. She forms the intention to push the Fatman and does so, regretting that her action will bring about the death of the Fatman, but thinking it for the best.

The death of the single man tied to the track and the death of the Fatman are both equally explained by Paul's and Pauline's quality of care respectively. Were it not for Paul's degree of care, he may well not have pulled the lever, and were it not for Pauline's degree of care, she may not have pushed the Fatman. In this case, both Paul and Pauline would seem to be open to accountable (and so of course attributably responsible as well) for their behaviours. There may of course be independent moral reasons for not pushing the Fatman, whereas there are no such reasons for not pulling

the lever, in which case Pauline would be more blameworthy than Paul.<sup>89</sup> Or some version of the Doctrine of Doing and Allowing could hold, in which case arguably Pauline would be more blameworthy. However, neither of these ways of discriminating between Paul and Pauline employs the intentional status of their behaviour.

Next consider the following, which is closer to the example of Truman's decision to order the use of the atomic weapons:

A paradigmatic application of DDE yields different assessments as to the moral permissibility for the respective action plans of two wartime pilots, one a strategic bomber (SB) and the other a terror bomber (TB). SB bombs a weapons cache as means to defeating the enemy while foreseeing that his bombing will bring about a number of civilian deaths. TB bombs the same number of civilians directly as a means to defeating the enemy (he intends to demoralize the enemy). (Delaney, 2008: 336)

Defenders of the DDE regularly argue<sup>90</sup> that the TB's action plan is morally impermissible as it involves an intention to bring about the civilian deaths, whereas the SB's action plan might be morally permissible as it does not entail an intention to bring about the civilian deaths, these deaths are merely foreseen. In the case of TB, the deaths are intended. What about in the case of SB? According to TIS, the deaths of the civilians are an intentional consequence of SB's intentional activity of bombing the weapons cache. The death of the civilians is a consequence of bombing the weapons cache, and SB both has the belief that bringing about the deaths of the civilians is, given her evidence, a possible part of the fulfilment of her aim of defeating the enemy by bombing the weapons cache and it is the case that the death of the civilians is brought about under sufficient guidance of System 2 Oversight and in line with the agent's plan to defeat the enemy. Here again we have an intended consequence and an intentional consequence.

---

<sup>89</sup> I take it that this is what is going on in cases like Transplant Surgeon. It seems plausible that surgeons have, as one of their operative characteristics in the role of doctors, the expectation that they will not intentionally kill, or not harm those who don't require it for treatment.

<sup>90</sup> See Delaney, 2007, 2008; Masek, 2010; and Sartorio, 2015, among others.

This case unfolds in the same way as that of Paul and Pauline. Both TB's and SB's behaviour are equally well explained, and in an equally normal way, by their quality of care. Both TB and SB are prepared to bring about the same number of deaths to meet their aims – both may regret this, or wish there were alternatives, but both are prepared to make the 'hard' choice. On the matter of alternatives, it should be noted that these do matter significantly. If it was the case that if SB had an alternative means to achieve his goal that did not involve the death of the civilians he would take it, but TB would not, then TB's behaviour is clearly explained by a greater lack of care than SB. But note that this works both ways. But it may be thought that because TB *intends* to kill the civilians he is less open to alternatives. However, it is important to remember that both TB and SB have the same aim – this is a necessary feature of a DDE case. This aim is to defeat the enemy, and if TB becomes aware of an alternative means to defeat the enemy that does not involve killing the civilians, he would be perfectly rational to change his plans.

Finally, it could be argued that the DDE can be defended by stressing (2) – that the means must be at least neutral. The argument will go that even though the intentional status itself might not be what captures the morally relevant difference, the moral valence of the means does. I find this unconvincing for two reasons: the first and most obvious is that my interest is purely in whether the outcome of a given activity being intended or intentional is morally relevant, all else equal. The second response is to question why the distinction between means and by-products should be thought of as morally relevant. It seems again that whether a means or a by-product are explained in a normal way by the agent's quality of care is not dependent on its being a means or a by-product. What an agent chooses as her means toward an end may be no more or less an expressive of her care than a by-product of this action.

In conclusion then, once the relationship between moral responsibility and intentional status is properly understood, the objection against Anscombe's protest that sought to diminish Truman's responsibility by denying positive intentional status to the outcome of his decision can be seen to be misguided and incorrect. If Truman is in fact off the hook, it is not thanks to any consideration of the intentional status of his behaviour.

## Concluding remarks

In this chapter I have sought to provide a partial account of moral responsibility sufficient to the task of allowing for the unravelling of the relationship between moral responsibility and the intentional status of activities and outcomes. This involved locating my account as one that was merit-based, rather than consequentialist, and explicating that my account can be treated as modular for the purposes of the free will debate.

After briefly clarifying how I understand the relationship between moral praiseworthiness, blameworthiness, and responsibility – which involved differentiating between the conditions for an agent to be open to moral responsibility and the conditions for an agent being praiseworthy or blameworthy – as well as exploring some of the complexities that arise when tackling the relationship between moral responsibility and intentional action, I proceeded to investigate the current discussion regarding the role of control in moral responsibility. This involved expository work on the volitionist and attributivist approaches to moral responsibility, since the former argues that there is necessary control condition on responsibility, while the latter do not. Of particular importance here was my consideration of the *explanatory quality of will* account which I took to be the most compelling attributivist account overall, and on which I build much of my own account.

Having explored these existing accounts, I then followed Watson and Shoemaker in pushing an objection applicable to accounts of both approaches: that no invariantist account of moral responsibility can capture the ambivalent nature of our responses to certain cases of moral responsibility. I also follow them in taking the best solution to this objection to be the endorsement of pluralism about moral responsibility. After briefly considering both Watson's and Shoemaker's attempts at pluralist accounts of moral responsibility, I introduce my own dual-variant account. This account builds on the excellent foundations of the explanatory quality of will account, but introduces a pluralist element into the account by exploiting an ambiguity in what it means to say that *a behaviour is explained in the normal way by an agent's quality of care*. The pluralism in question takes there to be two variants of moral responsibility: attributability and accountability. Crucially, the latter variant – which legitimates

more severe moral responses than the former, and so is open to a demand from fairness for control or avoidability – has a control condition, while the former does not. Given the central and constitutive role of control in the account of intentional activities and outcomes that I developed in Chapter 2, at this point I conclude that it is accountability, not attributability, that stands in a necessary relationship with intentional behaviour.

To prove that this is the case, I showed that the control condition on moral accountability is best understood as *normal* System 2 Oversight, a qualified version of the control condition on positive intentional status, where this requires that the System 2 oversight mechanism in question be functioning at least normally. Such a mechanism is taken to be functioning normally when it is functioning sufficiently in line with its operative characteristics, in what I called the *Normal Band*. It was shown that sufficient capacity to be responsive to moral reasons was part of System 2 oversight's operative characteristics, and that this explained why cases such as those of Skip and Harris are ones where the agent is not open to moral accountability despite acting intentionally – these agents lack the sufficient capacity to be responsive to moral reasons. I then extended this condition on accountability to intentional consequences, by similarly producing a qualified version of Transfer of intentional status, namely: Transfer of openness to accountability. Due to the relationship between the control condition on positive intentional status and moral accountability, I argued that the relationship between moral responsibility and the intentional status of a behaviour is best understood in the following way: an agent is open to accountability for a given behaviour only if that behaviour is intentional.

Finally, I closed out the chapter by returning to consider the motivating case, that of Anscombe and Truman. I argued that the defence of Truman that employed the DDE is unsuccessful, as there is no morally relevant distinction as far as responsibility is concerned (both variants) between an agent's intended outcomes and intentional outcomes solely based on whether they are intended or intentional.

## CONCLUSION

In the preceding three chapters I have set out to unravel the relationship between moral responsibility and intentional action. As has been shown, this relationship is best understood neither in terms of undifferentiated moral *responsibility* nor merely intentional *action*. Rather, I argued that the crucial relationship was between moral *accountability* and intentional *activities and outcomes*, and that this relationship is a necessary one. To do this, I set out to first provide a convincing account of intentional activity that featured a central control condition on positive intentional status: *System 2 Oversight*. I then turned to an investigation of moral responsibility and the eventual development of a pluralist account thereof, with one of the variants of responsibility I argue for – *accountability* – having a similar (but not identical) control condition to positive intentional status, namely: *normal System 2 Oversight*. In promulgating these two accounts, I demonstrated that the necessary relationship between accountability and positive intentional status is found in the *nexus of control*, by which I mean the fact that both notions share aspects of the same control condition: *System 2 Oversight*. This relationship can be explicated as: *if an agent is open to moral accountability based on some activity or outcome, this activity or outcome must necessarily have positive intentional status*. With this insight in hand, I returned to the original motivating case, that of Anscombe and Truman, and played out some of the consequences of my arguments for the Doctrine of Double Effect. I concluded that, at least so far as Truman's decision to use the atomic weapons goes, the fact that he did not *intend* to kill civilians but that this was only an *intentional* consequence has no inherent bearing on his openness to moral responsibility for this outcome – either in terms of the type or degree.

## BIBLIOGRAPHY

- Adams, F. 2006. Intentions Confer Intentionality upon Actions: A Reply to Knobe and Burra. *Journal of Cognition & Culture*. 6: 255-268.
- Adam, F. and Steadman, A. 2004. Intentional action and moral considerations: Still pragmatic. *Analysis*. 64: 268-276.
- Anscombe, G.E.M. 1957. Intention. *Proceedings of the Aristotelian Society*. 57: 321-332.
- Anscombe, G.E.M. 1963. *Intention* (second edition). United States of America: Harvard University Press.
- Anscombe, G.E.M. 1961. "War and Murder," in Stein, W. (ed.) 1961. *Nuclear Weapons: a Catholic Response*. London: Merlin: 43–62.
- Anscombe, G.E.M. 1974. "Practical Inference" in Geach, M. and Gormally, L. (eds.) 2005. *Human Life, Action and Ethics: Essays by G.E.M. Anscombe*. Imprint Academic.
- Anscombe, G.E.M. 1982. Medalist's Address: Action, Intention and 'Double Effect'. *Proceedings of the American Catholic Philosophical Association*. 56: 12-25.
- Aristotle. *Nicomachean Ethics*. Trans. Broadie, S. and Rowe, C. 2002. United Kingdom: Oxford University Press.
- Arpaly, N. and Schroeder, T. 1999. Praise, Blame and the Whole Self. *Philosophical Studies*. 93(2): 161- 188.
- Arpaly, N. and Schroeder, T. 2013. *In Praise of Desire*. Oxford Scholarship Online.
- Bayne, S.R. 2010. Elizabeth Anscombe's Intention. U.S.A.: Booksurge Publishing.

- Björnsson, G. 2017a. “Explaining (Away) the Epistemic Condition on Moral Responsibility” in Robichaud, P. and Wieland, J. W (eds.) 2017. *Responsibility: The Epistemic Condition*. Oxford Scholarship Online.
- Björnsson, G. 2017b. “Explaining Away Epistemic Skepticism about Culpability” in Shoemaker, D. (ed.) 2017. *Oxford Studies in Agency and Responsibility Volume 4*. Oxford Scholarship Online.
- Bratman, M. 1984. Two Faces of Intention. *The Philosophical Review*. 93(3): 375-405.
- Bratman, M. 1985. “Davidson's Theory of Intention” in *Faces of Intention*. 1999. Cambridge: Cambridge University Press.
- Bratman, M. 1987. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.
- Bratman, M. 2000. Fischer and Ravizza on Moral Responsibility and History. *Philosophy and Phenomenological Research*. 61(2): 453-458.
- Bratman, M. 2009a. “Intention, belief, and instrumental rationality” in Sobel, D. and Wall, S (eds.) 2009. *Reasons for Action*. Cambridge: Cambridge University Press.
- Bratman, M. 2009b. Intention, Practical Rationality, and Self-Governance. *Ethics*. 119(3): 411-443.
- Bratman, M. 2013. The Interplay of Intention and Reason. *Ethics*. 123(4): 657-672.
- Broome, J. 2013. *Rationality Through Reasoning*. UK: Wiley Blackwell.
- Bryne, R. W., Sanz, C. M., and Morgan, D. B. 2013. “Chimpanzees plan their tool use” in Crickette, M., Sanz, M., Call, J., and Boesch, C. (eds.) 2013. *Tool Use in Animals: Cognition and Ecology*. Cambridge University Press.

Carruthers, P. “The Fragmentation of Reasoning” in Quintanilla, P., Mantilla, C., and Cépeda, P. (eds.) 2014. *Cognición social y lenguaje. La intersubjetividad en la evolución de la especie y en el desarrollo del niño*. Lima: Pontificia Universidad Católica del Perú.

Crucius, C. 1744. “Guide to Rational Living” in Schneewind, J. B. (ed.), 1990. *Moral Philosophy from Montaigne to Kant*, vol.ii. Cambridge: Cambridge University Press.

Davidson, D. 1982. Rational Animals. *Dialectica*. 36: 317–328.

Davidson, D. 2001. *Essays on Actions and Events*. Oxford: Clarendon Press.

Delaney, N. F. 2007. A Note on Intention and the Doctrine of Double Effect. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*. 134(2): 103-110.

Delaney, N. F. 2008. Two Cheers for “Closeness”: Terror, Targeting and Double Effect. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*. 137(3): 335-367.

Doris, J. M., Knobe, J. and Woolfolk, R. L. 2007. Variantim about Responsibility. *Philosophical Perspectives*. 21: 183-214.

Driver, J. “Gertrude Elizabeth Margaret Anscombe” in Zalta, E.N. (ed.) 2011. *The Stanford Encyclopedia of Philosophy*. <<http://plato.stanford.edu/archives/win2011/entries/anscombe/>>.

Evans, J. S. B. T. 2009. “How many dual process theories do we need? One, two, or many?” in Evans, J. S. B. T. and Frankish, K (eds.) 2009. *In Two Minds: Dual Process and Beyond*. Oxford: Oxford University Press.

Evans, J. S. B. T. and Frankish, K (eds.) 2009. *In Two Minds: Dual Process and Beyond*. Oxford: Oxford University Press.

Evans, J. S. B.T., and Over, D. E. 1996. *Rationality and Reasoning (Essays in Cognitive Psychology)*. Taylor and Francis: Kindle Edition.

Farrer, C., Franck, N., Paillard, J., and Jeannerod, M. 2003. The role of proprioception in action recognition. *Consciousness and Cognition*. 12: 609-619.

Finkelstein, C. 2005. Responsibility for Unintended Consequences. *Ohio State Journal of Criminal Law*. 2: 579-599.

Fischer, J. M. 2004. Responsibility and Manipulation. *The Journal of Ethics*. 8: 145-177.

Fischer, J. M. 2007. “Compatibilism” in Fischer, J. M., Kane, R., Pereboom, D. and Vargas, M. 2007. *Four Views on Free Will*. Singapore: Blackwell Publishing.

Fischer, J. M. and Ravizza, M. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press.

Fodor, J. A. 1975. *The Language of Thought*. Cambridge/Massachusetts: Harvard University Press.

Ford, A. “Action and Generality” in Ford, A., Hornsby, J. and Stoutland, F. (eds.) 2011. *Essays on Anscombe's Intention*. Harvard University Press.

Frankfurt, H. 1969. Alternate Possibilities and Moral Responsibility. *Journal of Philosophy*. 66: 829-39.

Frankfurt, H. 1988. *The importance of what we care about*. Cambridge/New York/Melbourne: Cambridge University Press.

Gormally, L. “On Killing Human Beings” in Gormally, L., Jones, D. A. and Teichmann, R. (eds.) 2016. *The Moral Philosophy of Elizabeth Anscombe*. Exeter: Imprint Academic.

Grice, H.P. "Logic and Conversation" in Cole, P. and Morgan, J. (eds.) 1975. *Syntax and Semantics*, vol. 9.

Grice, H.P. 1971. Intention and Uncertainty. *Proceedings of the British Academy*. 5: 263–279.

Haggard, P. and Libet, B. 2001. Conscious Intention and Brain Activity. *Journal of Consciousness Studies*. 8(11): 47–63.

Haggard, P. and Tsarkis, M. 2009. The Experience of Agency: Feeling, Judgements, and Responsibility. *Association for Psychological Science*. 18(4): 242-246.

Haji, I. 2002. Compatibilist Views of Freedom and Responsibility in Kane, R. (ed.) 2009. *The Oxford Handbook of Free Will*.

Haldane, J. 2011. Identifying privative causes. *Analysis*. 71(4): 611-619.

Harman, G. 1976. "Practical Reasoning" in Mele, A. (ed.) 1997. *The Philosophy of Action*. Oxford: Oxford University Press.

Harman, G. 1986. "Willing and Intending" in Philosophical Grounds of Rationality, Grandy, R.E. and Warner R. (eds.) 1986. *Philosophical Grounds of Rationality*. Oxford: Oxford University Press.

Harman, G. 1999. *Reasoning, Meaning and Mind*. Oxford University Press.

Harman, G. 2006. Intending, Intention, Intent, Intentional Action, and Acting Intentionally: Comments on Knobe and Burra. *Journal of Cognition and Culture*. 6: 269-276.

Heuer, U. 2014. Intentions and the Reasons for which we Act. *Proceedings of the Aristotelian Society*. 114(3).

Holton, R. 2009. *Willing, Wanting, Waiting*. New York: Oxford University Press.

Holton, R. 2015. *In Praise of Desire*, by Arpaly, N. and Schroeder, T. Reviewed in: *Notre Dame Philosophical Reviews*.

Hornsby, J. 1980. *Actions*. London: Routledge.

Hornsby, J. 2012. “Actions and Activity” in ACTION THEORY, Sosa, E., Villanueva, E., and Brogaard, B. (eds.) 2012. *Philosophical Issues* 22. Wiley-Blackwell: Boston and Oxford.

Hursthouse, R. 1991. Arational Actions. *Journal of Philosophy*. 88(2): 57–68.

Kahneman, D. 2011. *Thinking, Fast and Slow*. Great Britain: Penguin Random House.

Kane, R. 1996. *The Significance of Free Will*. New York: Oxford University Press.

Kane, R. 2002. *Free will*. Oxford: Blackwell Publishers.

Kane, R. 2007. “Libertarianism” in Fischer, J. M., Kane, R., Pereboom, D. and Vargas, M. 2007. *Four Views on Free Will*. Singapore: Blackwell Publishing.

Killen, M. and de Waal, F. B. M. “The evolution and development of morality” in Aureli, F. and de Waal, F. B. M. (eds.) 2000. *Natural conflict resolution*. Berkeley: University of California Press.

King, M. 2014. Moral Responsibility and Merit. *Journal of Ethics & Social Philosophy*. 6(2): 1-17.

Knobe, J. 2003. Intentional action and side effects in ordinary language. *Analysis*. 63: 190-193.

Knobe, J. 2004. Intention, Intentional Action and Moral Considerations. *Analysis*. 64(2): 181-187.

Knobe, J. and Burra, A. 2006. The Folk Concepts of Intention and Intentional Action: A Cross-Cultural Study. *Journal of Cognition and Culture*. 6(1-2): 113-132.

Levy, N. 2005. The good, the bad and the blameworthy. *Journal of Ethics and Social Philosophy*. 1(2): 2-16.

Levy, N. 2008. Restoring Control: Comments on George Sher. *Philosophia*. 36: 213-221.

Levy, N. 2011. *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford Scholarship Online.

Levy, Y. 2017. Why cognitivism? *Canadian Journal of Philosophy*. 48(2): 223-244.

Libet, B., Gleason, C.A., Wright, E.W. and Pearl, D.K. 1983. Time of conscious intention to act in relation to onset of cerebral activity (readiness potential): The unconscious initiation of a freely voluntary act. *Brain*. 102: 623-642.

Locke, J. 1690. *An Essay Concerning Human Understanding, Volume 1*. ebook.

Mail&Guardian. 2013. “Oscar Pistorius to face premeditated murder charges in March.” <<http://mg.co.za/article/2013-08-19-oscar-pistorius-trial-set-for-march-3>> (Retrieved: 29 October 2013).

Malle, B. F. and Knobe, J. 1997. The folk concept of intentionality. *Journal of Experimental Social Psychology*. 33(2). 101-121.

Mangan, J. T. 1949. An historical analysis of the principle of double effect. *Theological Studies*. 10(1): 41-61.

Masek, L. 2010. Intentions, Motives and the Doctrine of Double Effect. *The Philosophical Quarterly*. 60(240): 567-585.

Mason, E. 2005. *Moral Responsibility*. Oxford: Blackwell Publishing.

McCann, H. J. 1991. Settled objectives and rational constraints. *American Philosophical Quarterly*. 28: 24-36.

McCann, H. J. 2005. Intentional action and intending: Recent empirical studies. *Philosophical Psychology*. 18(6): 737-748.

McIntyre, A. 2001. Doing Away with Double Effect. *Ethics*. 111(2): 219-225.

McKenna, M. 1997. Alternative Possibilities and the Failure of the Counterexample Strategy. *Journal of Social Philosophy*. 28: 71-85.

McKenna, M. 2003. "Robustness, Control, and the Demand for Morally Significant Alternatives" in Widerker and McKenna (eds.) 2003. *Moral Responsibility and Alternative Possibilities*. Aldershot: Ashgate Press.

McKenna, M. 2008. Compatibilism & Desert: Critical Comments on "Four Views on Free Will". *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*. 144(1): 3-13.

McKenna, M. 2009. "Compatibilism" in Zalta, E. N. (ed.) 2009. The Stanford Encyclopedia of Philosophy. <<http://plato.stanford.edu/archives/win2009/entries/compatibilism/>>.

McKenna, M. 2012. "Directed Blame and Conversation" in Coates, J. D. and Tognazzini, N. A. (eds.) 2012. *Blame: Its Nature and Norms*. Oxford Scholarship Online.

Mele, A. R. and Moser, P. K. 1994. Intentional Action. *Noûs*. 28(1): 39-68.

Mele, A. R. 2006. Fischer and Ravizza on Moral Responsibility. *The Journal of Ethics*. 10(3): 283-294.

Mele, A. R. 2009. *Effective Intentions: The Power of Conscious Will*. Oxford Scholarship Online.

Metcalfe, J. and Greene, M. J. 2007. Metacognition of agency. *Journal of Experimental Psychology: General*. 136(2): 189-199.

- Millican, P. and Wooldridge, M. 2014. “Them and Us: Autonomous Agents In Vivo and In Silico” in Baltag, A. and Smets, S. (eds.) 2014. *Johan van Benthem on Logic and Information Dynamics (Outstanding Contributions to Logic Volume 5)*. Springer-Verlag.
- Moors, A. and De Houwer, J. 2007. “What is automaticity: An analysis of its component features and their interrelations” in Bargh, J. A. (Ed.) *Social psychology and the unconscious: The automaticity of higher mental processes*. New York: Psychology Press.
- Nado, J. 2008. Effects of Moral Cognition on Judgements of Intentionality. *British Journal of the Philosophy of Science*. 59: 709-731.
- Norman, D. A. and Shallice, T. “Attention to Action” in Davidson, R. J., Schwartz, G. E., and Shapiro, D. (eds.) 1986. *Consciousness and Self-Regulation*. Springer US.
- O’Shaughnessy, B. 1973. Trying (As the Mental “Pineal Gland”). *The Journal of Philosophy*. 70(13): 365-386.
- Paul, S. 2013. “Intention” in LaFollete, H. (ed.) 2013. *The International Encyclopedia of Ethics*. Blackwell Publishing Ltd.
- Pereboom, D. 2007. “Hard Incompatibilism” in Fischer, J. M., Kane, R., Pereboom, D. and Vargas, M. 2007. *Four Views on Free Will*. Singapore: Blackwell Publishing.
- Pereboom, D. 2009. Hard Incompatibilism and Its Rivals. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*. 144(1): 21-33.
- Pereboom, D. 2013. “Free Will Skepticism, Blame, and Obligation” in Tognazzini, N. and Coates, D. J. (eds.) 2013. *Blame: Its Nature and Norms*. New York: Oxford University Press.
- Plotnik, J.M., de Waal, F. B. M., and Reiss, D. 2006. Self-recognition in an Asian elephant. *PNAS*. 103: 17053–17057.

Quinn, W. S. 1989. Actions, Intentions, and Consequences: The Doctrine of Double Effect. *Philosophy and Public Affairs*. 18(4): 334-351.

Rabinowicz, W. and Ronnow-Rasmussen, T. 2004. The Strike of the Demon: On Fitting Pro-attitudes and Value. *Ethics*. 114: 391-423.

Radoilska, L. 2016. Rethinking Responsibility: The Role of Judgement and Belief. Paper presented at *Rethinking Responsibility* at Birkbeck College, London.

Rosen, G. 2003. Culpability and Ignorance. *Proceedings of the Aristotelian Society*. New Series, 103: 61-84.

Rosen, G. 2014. Culpability, Duress and Excuses. *Proceedings of the Aristotelian Society*. 88: 69- 90

Ross, J. 2009. How to be a Cognitivist about Practical Reason. *Oxford Studies in Metaethics*. 4: 243–281.

Roughley, N. 2004. “Naturalism and Expressivism On the ‘Natural’ Stuff of Moral Normativity and Problems with its ‘Naturalisation’” in Schaber, P. (ed.) 2004. *Normativity and Naturalism*. Frankfurt/New York: Ontos.

SABC. 2013. “Pinetown truck driver faces 22 murder charges.” <<http://www.sabc.co.za/news/a/acc97b8041086672a14fa1434f2981a1/Pinetown-truck-driver-faces-22-murder-charges-20130909>> (Retrieved: 29 October 2013).

Sartorio, C. 2010. “Causation and Ethics” in Beebee, H., Hitchcock, C., and Menzies, P. (eds.) 2010. *The Oxford Handbook of Causation*. Oxford Handbooks Online.

Sartorio, C. 2015. “The Problem of Determinism and Free Will is Not the Problem of Determinism and Free Will” in Mele, A. (ed.) 2015. *Surrounding Free Will*. New York: Oxford University Press.

Sartorio, C. 2016. A Partial Defense of the Actual-Sequence Model of Freedom. *Ethics*. 20: 107-120.

- Sato, A, and Yasuda, A. 2005. Illusion of sense of self-agency: discrepancy between the predicted and actual sensory consequences of actions modulates the sense of self-agency, but not the sense of self-ownership. *Cognition*. 94(3): 241-255.
- Scanlon, T. M. 1986. The significance of choice. *The Tanner Lectures on Human Values*. 7: 149–216.
- Scanlon, T. M. 2008. *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge: Harvard University Press.
- Scanlon, T. M. 2015. “Forms and Conditions of Responsibility” in Clarke, R., McKenna, M., and Smith, A. M. (eds.) 2015. *The Nature of Moral Responsibility: New Essays*. Oxford Scholarship Online.
- Schlosser, M. E. 2013. Conscious Will, Reason-Responsiveness, and Moral Responsibility. *Ethics*. 17: 205–232.
- Searle, J. 1980. The Intentionality of Intention and Action. *Cognitive Science*. 1(4): 47-70.
- Searle, J. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.
- Seebaß, G., Schmitz, M., and Gollwitzer, P. M. (eds.) 2013. *Acting Intentionally and its Limits: Individuals, Groups, Institutions*. De Gruyter.
- Setiya, K. 2003. Explaining Action. *The Philosophical Review*. 112(3): 339-393.
- Setiya, K. 2007. Cognitivism about Instrumental Reason. *Ethics*. 117: 649–673.
- Setiya, K. 2008. Practical Knowledge. *Ethics*. 118(3): 388-409.
- Setiya, K. 2011. “Knowledge of Intention” in Ford, A., Hornsby, J. and Stoutland, F. (eds.) 2011. *Essays on Anscombe's Intention*. Harvard University Press.

- Setiya, K. 2011. “Intention” in Zalta, E.N. (ed.) 2011. *The Stanford Encyclopedia of Philosophy*. <<http://plato.stanford.edu/archives/spr2011/entries/intention/>>.
- Setiya, K. 2012. Knowing How. *Proceedings of the Aristotelian Society*. 112(3): 285-307.
- Shoemaker, D. 2015a. *Responsibility from the Margins*. Oxford Scholarship Online.
- Shoemaker, D. 2015b. “Ecumenical Attributability” in Clarke, R., McKenna, M., and Smith, A. M. (eds.) 2015. *The Nature of Moral Responsibility: New Essays*. Oxford Scholarship Online.
- Sliwa, P. 2017. “On Knowing What’s Right and Being Responsible for It” in Robichaud, P. and Wieland, J. W. (eds.) 2017. *Responsibility: The Epistemic Condition*. Oxford Online Scholarship.
- Smith, H. 1983. Culpable Ignorance. *The Philosophical Review*. 92: 543-571.
- Smith, M. 1994. *The Moral Problem*. Cornwall: Blackwell Publishing Ltd.
- Smith, M. 2011. “Scanlon on Desire and the Explanation of Action” in Freeman, S., Kumar, R., and Wallace, R. J. (eds.) 2011. *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon*. New York: Oxford University Press.
- Smith, M. 2012. Four Objections to the Standard Story of Action (and Four Replies). *Philosophical Issues*. 22, Action Theory: 387-401.
- Smith, M. 2010. What Else Justification Could Be. *Noûs*. 44(1): 10-31.
- Stanovich, K. E. “Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory?” in Evans, J. S. B. T. and Frankish, K (eds.) 2009. *In Two Minds: Dual Process and Beyond*. Oxford: Oxford University Press.
- Steward, H. 2000. Do Actions Occur Inside the Body?. *Mind & Society*. 1(2): 107-125.

Steward, H. 2012. ACTIONS AS PROCESSES. *Philosophical Perspectives*. 26: 373-388.

Stoutland, F. 2011. “Introduction: Anscombe’s Intention in Context” in Ford, A., Hornsby, J., and Stoutland, F. (eds.) 2011. *Essays on Anscombe’s “Intention”*. Harvard University Press.

Strawson, H. F. 1962. Freedom and Resentment. *Proceedings of the British Academy*. 48: 187-211.

Strawson, G. 1994. The Impossibility of Moral Responsibility. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*. 75(1/2): 5-24.

Sun, R., Lane, S. M., and Matthews, R. C. “The two systems of learning: An architectural perspective” in Evans, J. S. B. T. and Frankish, K (eds.) 2009. *In Two Minds: Dual Process and Beyond*. Oxford: Oxford University Press.

The New York Times. 2013. “‘Ugly Thoughts’ Defense Fails as Officer Is Convicted in Cannibal Plot.” <<http://www.nytimes.com/2013/03/13/nyregion/gilberto-valle-is-found-guilty-in-cannibal-case.html>> (Retrieved: 29 October 2013).

Thompson, M. 2008. *Life and Action*. Cambridge: Harvard University Press.

Velleman, J. D. 1989. *Practical Reflection*. Princeton: Princeton University Press.

Velleman, J. D. 1991. Intention, Plans, and Practical Reason by Michael E. Bratman. *The Philosophical Review*. 10(2): 277-288.

Velleman, J. D. 2014. *The Possibility of Practical Reason*. Ann Arbor: Michigan Publishing. Online Access: <<http://dx.doi.org/10.3998/maize.13240734.0001.001>>

Wallace, R. J. 1994. *Responsibility and the Moral Sentiments*. London: Harvard University Press.

- Wallace, R. J. 2009. “Practical Reason” in Zalta, E.N. (ed.) 2009. *The Stanford Encyclopedia of Philosophy*.  
<<http://plato.stanford.edu/archives/sum2009/entries/practical-reason/>>.
- Wallace, R. J. 2006. *Normativity and the Will*. Oxford: Oxford University Press.
- Watson, G. 2004. *Agency and Answerability*. Oxford: Oxford University Press.
- Williams, B. A. O. 1976. Moral Luck. *Proceedings of the Aristotelian Society, Supplementary Volumes*. 50: 115–35.
- Wilson, G. and Shpall, S. “Action” in Zalta, E.N. (ed.) 2012. *The Stanford Encyclopedia of Philosophy*.  
<<http://plato.stanford.edu/archives/sum2012/entries/action/>>.
- Wolf, S. 1990. *Freedom within Reason*. Oxford and New York: Oxford University Press.
- Woolfolk, R. L., Doris, J. M. and Darley, J. M. 2006. Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility. *Cognition*. 100: 283–301.
- Woollard, F. 2012a. The Doctrine of Doing and Allowing I: Analysis of the Doing/Allowing Distinction. *Philosophy Compass*. 7(7): 448-458.
- Woollard, F. 2012b. The Doctrine of Doing and Allowing II: The Moral Relevance of the Doing/Allowing Distinction. *Philosophy Compass*. 7(7): 459-469.
- Woollard, F. 2017. Double effect, doing and allowing, and the relaxed nonconsequentialist. *Philosophical Explorations*. 20(2): 142-158.