# Reasons-Responsiveness, Action and Control:

# An Event-Causal Account of Agency

Jingbo Hu

Summitted for the Degree of Doctor of Philosophy

University of Sheffield

November 2020

## Acknowledgements

First of all, I would like to thank Jules Holroyd and James Lenman, my two supervisors, who have provided me with a lot of academic and emotional support during my PhD.

In addition, with respect to writing this thesis, I want to thank many other people who have provided me with comments, advice or other assistance. They are Luca Barlassina, Yonatan Shemmer, William Hornett, Alex Duval, Stephen Laurence, Matthew Cull, James Lewis, Brendan Kelters, Andreas Bunge, Radivoj Stupar, Pekka Väyrynen, Derek Baker, Jessica Wilson, Timothy Perrine, Alexander Gebharter, Qianqian Sun, Junwei Wang, Qiaoyin Lu, Helen Beebee, T. Ryan Byerly, Andrei Buckareff, Kenneth Silver, Andrew Christman, Ju Chen, Mu Liu, Shaun Nichols, Gunnar Björnsson and Ruoyu Zhang.

I also want to thank my friends at Sheffield, with whom I have had wonderful time in hiking at Peak District, playing board games, drinking, having hotpot and sharing many other nice foods. They are Yifan Mei, Felicity Fu, Cheer Cheng, Can Du, Jack Herbert, Victoria Kononova, María José Pietrini, Toshiaki Ito, Barend de Rooij, Kayleigh Doherty, Dongdong Yang and Giulia Casini.

Finally, I will express my gratitude to my family, especially my mother, Lingyou Lv, and my father, Guojie Hu. My pursuit of philosophy is conditioned on their unconditional love towards me.

Research for this thesis was funded by the University of Sheffield-CSC Joint Scholarship (2016) and the Great Britain-China Educational Trust (2019).

## Abstract

In this thesis, I aim to contribute to the reconciliation of two ways of looking at human agency—from the perspective of agents themselves, and from a detached, scientific perspective—by combining resources from the free will literature and the action theory literature. I will show that we can preserve most of our ordinary conception and intuitions about human agency rooted in common sense even if we suppose the truth of determinism and a universal event-causal framework. Below are the two key claims defended in the thesis.

> (i.) Free Agency (which is required by moral responsibility) is the ability to respond to reasons.
>
> (ii.) To exercise free agency is to act for reasons; while acting for reasons can be captured within an event-causal framework.

Accordingly, I will defend two theories concerning human agency: one is the reasons-responsiveness theory of free will and moral responsibility; the other is the causal theory of action. The first theory helps to show that free will and moral responsibility can exist in a deterministic world; while the second theory helps to resolve the tension between action and the event-causal framework. Moreover, I will integrate these two theories. This integration aims to provide a more complete picture of human agency. According to this picture, the exercising of human agency can be taken as a continuum—ranging from basic voluntary control to high-level actions which merit moral evaluation; this continuum can be captured within a unified theoretical framework, namely, the event-causal framework. That is, the important features of human agency, from basic purposive actions to free and responsible actions can be explained in terms of event-causations.

In doing so, I advance a new account of reasons-responsiveness theory, deflate the problem of causal deviance by developing a novel account of control, and solve the problem of disappearing agency by developing an account of agent's participation in her action. I will also investigate several kinds of phenomenology of agency which purportedly speak against the event-causal account of action. The integrated theory developed shows us how our agency, in all its complexity, fits into a naturalistic picture of the world.

# Contents

**Chapter 1: Introduction……1**

The Two Images of Human Agency…….1

Reasons-Responsiveness Theories……7

Integrating the Causal Theory of Action……..17

The Agenda of the Thesis……23

**Chapter 2: Reasons-Responsiveness and The Frankfurt-Style Cases: A Middle Path……..26**

Introduction…….26

Reasons-Responsiveness as Conditional Freedom to Do Otherwise……..29

Fischer and Ravizza's Mechanism-Based Reasons-Responsiveness Account……37

Sartorio's Causal Reasons-Sensitivity Account……...45

How to Deal with the Frankfurt-Style Cases……..…53

Conclusion……..…59

**Chapter 3: Reasons-Responsiveness and Explanation……..…61**

Introduction…….....61

The Frankfurt-Style Cases and Reasons-Responsiveness……..….63

Sartorio's Argument for the Purported Inconsistency……....66

Reasons-Responsiveness and Explanation……..…72

Defending the Modal View of Reasons-Responsiveness……...78

Concluding Remarks: The Revenge of Compatibilism…….….83

**Chapter 4: The Problem of Causal Deviance……85**

Introduction……85

Some Clarifications for Causal Deviance……..…87

The Sophisticated Version of CTA……92

Davidson's Deflationism of CTA……..98

Control and Causal Deviance……101

Conclusion……113

**Chapter 5: Disappearing Agency and Two Intuitions about Action…….114**

Introduction……..114

Reductive Explanation and Agent-Participation……..116

Anti-Reduction……..132

Conclusion……137

**Chapter 6: Disappearing Agency and The Phenomenology of Agency…….138**

Introduction…….…..138

Disappearing Agency: A Phenomenological Reconstruction……139

The Phenomenology of Acting……..144

The Phenomenology of Making Choices and the Sense of Indetermination……...……..150

The Phenomenology of Exerting Effort……..…159

Conclusion……..161

**Conclusion…….162**

**Bibliography……..165**

# Chapter 1: Introduction

## 1. The Two Images of Human Agency

Wilfred Sellars once expressed his views on the aim of philosophy in the following ways:

> "The aim of philosophy, abstractly formulated, is to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term." (Sellars 1962/2007, 369)

According to Sellars, our current understanding about the world is shaped by two different images— *the manifest image* and *the scientific image*. Roughly speaking, the manifest image is constituted with everyday objects that can be directly observed and interacted with by us. To name a few, chairs, dogs and stones, etc. Apart from these concrete things, the manifest image also includes properties and abstract beings that we can unreflectively talk about in our daily discourse. These include colours, beauty, friendship and so on. The scientific image is what our best contemporary scientific theories present to us. The world is composed of particles that cannot be directly seen, such as molecules, atoms, protons, electrons and so on. The scientific image is apparently different from the manifest one. Most of the entities postulated and studied by science are not directly perceived by us. And most of the everyday objects are not directly dealt with by science. Accordingly, these two images are purportedly clashing with one another: ordinary concepts such as friendship and beauty, which are abstract and do not locate in space-time, are not easily made sense of by contemporary sciences. However, these two images are equally important to us. We cannot live easily only with one and without the other. Abandoning either image seems to be an imprudent decision to make. For Sellars, then, the primary aim of contemporary philosophy is to reconcile these two images.

I am not sure whether Sellars's conception of philosophy applies to all areas of philosophy. However, I think it accurately captures the main theme in philosophy of action—reconciling the clash of the two different images of human agency. Our agency is presented in quite different ways within two different images—the manifest image and the scientific image. The manifest image of agency is what our common sense tells us about agency. Within this picture, we have certain causal power in virtue of our agency. We make a difference to ourselves and on the world by devising plans, making decisions and acting. For example, if I get thirsty, I will get a glass of water to drink, I cause a series of events to happen by exercising my agency: a cup being raised, the tap being turned on and the water flowing away from the tap, and so on. In the manifest image, humans' actions frequently enter the causal network constituted

with other events. However, humans' exercising agency stands out from normal events happening in the world. Here is a list of characteristics which may distinguish our agency from other phenomena of the world:

> **Control**: we are exercising control over our bodies and the extension of our bodies (e.g., tools, objects) when we are acting. This characteristic of human agency is sometimes referred to as self-control or self-determination.
>
> **Reasons:** Normally, when we are acting, we are acting for specific reasons. And when we are, this explains our and others' actions with reference to reasons.
>
> **Choice:** Normally, when we are going to act, we are confronting a bunch of options, a bunch of alternative courses of action.
>
> **Moral Responsibility**: We certainly suppose that we, and others, are responsible for what we do.

To summarize, in the manifest image, human agency is special and distinguished from other occurrences of the world. The scientific image of agency, however, tells another story. As living organisms, humans are no more than a set of biological and neural systems. This implies that human agency is no exception to the world. Rather, human agency can be studied and analysed through natural sciences, just like any other beings (whether animate or inanimate) can be done with sciences.

The manifest image and the scientific image provide us with different lenses through which to view human agency. Within the manifest image, agency is understood through agents' motivations, choices and reasons, while in the scientific image it is understood through causes, laws of nature and mechanisms. Given these differences, there is the question whether the two images of human agency fit into one another and how they can reconcile. Specifically, there are two main obstacles to a reconciliatory project. The first is how it is possible to have free and responsible agency if determinism is true. The second is how it is possible to have action or agentive control over our bodies if the universal event-causation is true. In what follows I will present these two challenges in turn.

### 1.1 Free Will and Determinism

It is imprudent to simply judge that the scepticism about free will is due to the dissimilarities between the manifest image and the scientific image. We have a scientific image only after the birth of modern science. The free will problem, however, is perennial. It can be traced back to the ancient time, long before the birth of modern science. A more plausible diagnosis is that the free will problem arises because

we can reflect on our own agency from two different perspectives—the subjective perspective and the objective perspective. From the subjective perspective, we are free—we can make choices from a wide range of options; we can act following our interests, preferences and reasons. Moreover, since the concept of free will has a close relation to the concept of moral responsibility, from the subjective perspective, we hold ourselves morally responsible for what we do. By comparison, we can reflect on our choices and actions from a more objective perspective under which actions are influenced and motivated by factors beyond our control. For example, it is intuitive to think that our actions and choices flow from our personalities and characters; while the personalities and characters are largely shaped by the environmental factors, such as where we happen to live and how we are educated, factors over which we barely have control. The problem of free will become pressing when switching from one perspective to another. By doing this, we begin to question whether things that we do are really up to us.

So how does the tension between the subjective and objective perspective related to the tension between the two images of agency that just mentioned? Agency viewed from the subjective perspective is roughly equivalent to how agency is presented in the manifest image; while this objective perspective cannot directly be identified with the scientific image. As said, the objective perspective is more primitive—one can have an objective perspective of her agency without knowing anything about modern sciences. Nevertheless, the objective perspective is further backed up by the success of modern science and eventually evolves into the scientific image. Within the scientific image, actions are not only generated by factors beyond our control, they are also phenomena to be explained and studied by sciences. For actions are strictly governed by the laws of nature which equally apply to other animals or even objects. This is why the problem of the compatibility between free will and determinism occupies the central place in the free will discussion.

If our universe is strictly governed by laws of nature and that every event is strictly determined at the beginning of the universe, then how can we be genuinely free? [1] This is the threat from determinism, and many philosophers worry that it would rule out the existence of free will. Perhaps the most prominent argument to establish such a conclusion is the consequent argument.[2] The intuitions behind the argument can be articulated as follows:

i) we cannot change either the past or the laws of nature;

[1] By saying actions are strictly governed by the laws, I am here presuming the necessitarian view of laws of nature. (Armstrong 1983) The is not the only view. The Humean conception of the laws of nature has the potential to block the consequence argument.(Beebee and Mele 2002) I choose to present necessitarian view for it will make the determinism challenge more compelling.

[2] For a most sophisticated presentation and defence of this argument, see van Inwagen 1983.

ii) The past and the laws of nature together entail any events in the future (Determinism);

iii) We cannot change any events that happen in the future.

This argument directly applies to what we choose to do if actions are also covered by the laws of nature. This entails that we cannot do otherwise than what it is determined by the laws plus the remote past. For a long time, the compatibility of determinism and free will has been the central issue of the free will debate. The answers to the question of whether free will is compatible with determinism roughly divide philosophers of this topic into two camps: incompatibilism and compatibilism—incompatibilists think that determinism is incompatible while compatibilists think they are compatible.

Some may wonder why we should care about the compatibility of causal determinism and free will, given that the picture provided by contemporary science (quantum mechanics, in particular) is probably indeterministic. However, there are still some reasons to concern about whether determinism and free will are compatible.

First, even if the micro-world abides by indeterministic laws, it is possible that the macro-world is by and large deterministic. That is to say, human's actions and choice may still be governed by psycho-physiological laws which are deterministic. Human's free will may still be threatened by a kind of 'local determinism'.[3]

Second, the current understanding of quantum mechanics is not final and perfect. There are different but empirically equivalent interpretations for quantum mechanics. One of them is Bohmian mechanics, which provides a deterministic description of the quantum phenomena. Thus, before physicists discover the fundamental theory, we are not sure whether the universe is deterministic or not.[4]

Third, we can get a better understanding of the conditions for free will by studying its relation to determinism. Note that even indeterminism alone cannot guarantee the existence of free will. If humans' choices and actions come about through an indeterministic process, then it seems to follow that these choices and action just take place by luck. Randomness would equally undermine human's control over their choices and actions. Thus, it is not clear whether free will can be compatible with indeterminism either. Fortunately, the strategy to deal with determinism may enable us to deal with indeterminism as well. For example, if we can show that the truth of determinism is *irrelevant* to the existence of free will and moral responsibility, as many philosophers have tried to do, then we need to worry about the threat from indeterminism.

[3] See for example, Honderich 1988; Kane 2005.
[4] For this point, see Hoefer 2016.

These reasons explain why the compatibility of free will and determinism is still relevant to the reconciliation of the two images of human agency. Now I turn to another issue which also occupies the frontline of the reconciliatory project.

### 1.2 The Nature of Action and Event-Causation

Apart from the debate in free will, another central concern in philosophy of action is the nature of action. At first appearance, these questions seem not to have much to do with the reconciliatory project mentioned above. Rather, they are more like other traditional philosophical projects which aim to provide conceptual analysis for certain concepts. However, if we look closely into the contexts in which these questions are being addressed, we will find the main tenet of these questions are also about reconciling the manifest picture and the scientific picture.

As Wittgenstein (1953/2009, 169) famously raises the question "what is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?", the nature of action intrigues many philosophers because actions differ so much from mere bodily movements such as trembling, coughing and sneezing. Actions are what we perform spontaneously and intentionally, while the bodily movements such as trembling and knee-jerk reflex are events that just happen to us. More importantly, actions are usually done for *reasons*. We can evaluate or justify our actions from a *rational perspective*. This ordinary conception of action, however, is difficult to fit into the picture revealed by sciences, according to which all the happenings in the world are just causal interactions of events.[5] It is difficult to imagine that action can arise merely from such a picture. For example, Erasmus Mayr writes:

> …when we start to reflect on the nature of those actions and about our supposedly active role in performing them, we begin to be puzzled by the question of how the picture of us as active beings can be reconciled with another picture of the world to which we have grown accustomed since the eighteenth century—an image of the world as a flux of events, following upon each other, where one event can be explained by appeal to prior events and to natural laws discovered by natural science, if it can be explained at all. As we are part of the world, it seems that we must also fit ourselves in this latter, 'scientific',

[5] There are alternative accounts such that the fundamental causal relations are substance-causation (e.g., Lowe 2008); or that event-causation and substance-causation are both fundamentally required by our ontology (e.g., Steward 2012; Hyman 2015). In this thesis, I will suppose that the scientific picture reveals an event-causation framework for this is the mainstream idea.

> picture of the world—and this latter picture seems to have no room for activity, but only for happenings that befall the inhabitants of this world. (Mayr 2011, 1)

The nature of action is relevant to the reconciliatory project because the features of actions in the manifest image cannot find their counterparts in the scientific image, according to which the only causal interactions in the natural world (or the world revealed by natural sciences) are event-causation. Intuitively, causal interactions among events cannot add up to anything like exercising control and acting for reasons, which are crucial aspects of action. Just like the deterministic picture leads to scepticism about the ability to acting freely, the event-causal framework leads to scepticism about the very ability to act. This scepticism of action has once been presented by a character in Pynchon's novel:

> "But you had taken on a greater, and more harmful, illusion. The illusion of control. That A could do B. but that was false. Completely. No one can *do*. Things only happen." (Pynchon 1973, 34, requoted from Dennett 2004, 26) [6]

In philosophy of action, the discussion about free will is usually independent of the discussion about the nature of action. However, these two areas can and should be unified by a single project. That is, the project to reconcile the tension between two different images of human agency: from the manifest image, we view human agency as special and exceptional—humans act for reasons and usually act in a free and responsible way; from the scientific image, human agency is part of the event-causal nexus and covered by laws of nature. Specifically, the scientific picture challenges our ordinary conception of action from two respects: determinism challenges the idea that we are free agent, while the event-causal framework challenges the idea that we are capable of acting.

### 1.3 The Project of the Thesis

In this thesis, I aim to contribute to the reconciliation of two images of human agency by combining resources from the free will literature and the action theory literature. I will show that we can reserve most of our ordinary conception and intuitions about human agency rooted in the manifest image even if we suppose the truth of determinism and a universal event-causal framework. Below are the two key claims defended in the thesis.

[6] Philosophers like Bishop (1989, 1-2) and Steward (2012, 2) have similar comments that the naturalistic picture has the danger of leading to scepticism about general agency.

(i.) Free Agency (which is required by moral responsibility) is the ability to respond to reasons.

(ii.) To exercise free agency is to act for reasons; while acting for reasons can be captured within an event-causal framework.

Accordingly, I will defend two theories concerning human agency: one is the reasons-responsiveness theory for free will and moral responsibility; the other is the causal theory of action. The first theory helps to show that free will and moral responsibility can exist in a deterministic world; while the second theory helps to resolve the tension between action and the event-causal framework. Moreover, I will integrate these two theories. This integration aims to provide a more complete picture of human agency. According to this picture, the exercising human agency can be taken as a continuum—ranging from basic voluntary control to high-level actions which merit moral evaluation; this continuum can be captured within a unified theoretical framework, namely, the event-causal framework. That is, the important features of human agency, from basic purposive actions to free and responsible actions can be explained in terms of event-causation and counterfactual. In what follows, I will set up some essential backgrounds and the motivations for the project, then outline the structure of the arguments of the thesis.

## 2. Reasons-Responsiveness Theory

### 2.1 An Outline of the theory

The reasons-responsiveness theory is one of the most popular compatibilist theories in the literature. In a narrower sense, the 'reasons-responsiveness theory' is reserved for the theory developed by Fischer (and sometimes with Ravizza) and the followers. In this thesis, I will use the term in a broad sense: it refers to a bunch of theories proposed by different philosophers (e.g., Brink & Nelkin 2013; Fischer 1994, 2012; Fischer & Ravizza 1998; McKenna 2013; Wolf 1990). All these theories share a key idea: humans' freedom required by moral responsibility lie in their capacity to act in accordance with good reasons.[7] To illustrate, consider the following pair of scenarios.

Chun is now penniless and hungry. He has no better way to get food so he decides to steal some from the supermarket. But in making his decision, Chun is reasons-responsive in the sense that: If Chun knew other ways to get food, e.g., from some charity institutions, he would refrain from making the decision; If Chun knew that the supermarket was under good surveillance and he would have a high chance of being

[7] Apart from reasons-responsiveness, philosophers use different terms to denote the ability to respond to reasons. For example, 'reflective self-control' (Wallace 1996), and 'rational ability' (Nelkin 2011).

caught, he would refrain from making the decision. Now consider another case, Zhong, who is not in short of money or food but suffers from Kleptomania (which means that he is addicted to stealing) decides to steal some food from the supermarket. In making his decision, Zhong is not reasons-responsiveness in the sense that no matter what sufficient reasons tell him not to steal, he cannot help stealing because of the mental illness. We have different attitudes towards Chun and Zhong's behaviour. We think that Chun is morally responsible for what he does but while Zhong is not. And the reasons-responsiveness has an explanation, that is, Chun is reasons-responsive while Zhong is not.

In this thesis, I will defend a leeway approach to reasons-responsiveness, according to which, i) reasons-responsiveness consists of the conditional ability to do otherwise, to which I refer as the agentive ability conception;[8] ii) the notion of reasons-responsiveness is analysed in conditionals, to which I refer as the modal conception. Very roughly, according to the leeway approach, an agent is reasons-responsive, iff, if there are sufficient reasons to do otherwise, then the agent would do otherwise.[9]

The leeway approach provides a straightforward specification for the notion of reasons-responsiveness. Besides, this approach has an obvious advantage: regarded as a conditional ability to do otherwise, reasons-responsiveness is compatible with determinism. Determinism is threatening to free will because it seemingly rules out people's ability to do otherwise. Thus, employing a conditional analysis to handle the challenge from determinism was once a popular strategy from classical compatibilism (e.g., Hume 1748/2008; Ayer 1954; Davidson 1973). The idea is that that notion of the 'can do otherwise' is analysed in a conditional way such that if the agent has a different mental antecedent (e.g., chooses, desires or intends to do otherwise), he can do otherwise. Analysed in this way, the freedom to do otherwise is compatible with determinism. For all it requires is that if the past is different (say, the agent has a different mental antecedent), then the future will be different (say, the agent act differently from the actual situation). This does not contradict the thesis of determinism. However, there is a well-known problem with classical compatibilism—it qualifies agents suffering from addiction and mental disorders as free agents, which is counter-intuitive. The leeway approach to reasons-responsiveness can solve this problem for it identifies freedom as an ability to respond to reasons rather than an ability to respond to mental antecedents such as choices and desires. As I will show in Chapter 2, the reasons-responsiveness theory can be viewed as a revision of classical compatibilism.

[8] This, of course, differs from the mainstream reasons-responsiveness theory defended by Fischer and Ravizza (1998), according to which, reasons-responsiveness does not amount to the ability to do otherwise; rather, it is a modal property of the mechanism leading to the action.

[9] A related question is what degrees of reasons-responsiveness is sufficient to ground freedom and moral responsibility. I will leave this question open in my thesis.

Setting aside the leeway conception, there are still several important motivations for a general reasons-responsiveness theory. First, we have a deep-rooted conviction that our special moral status is closely connected to our rational faculties. This conviction is also recorded in a prominent philosophical tradition which relates human's responsible agency with his ability to regulate himself with reasons. For example, this idea can trace back to Aristotle, who places great weight on human's deliberation and the ability to make decisions based on the 'ultimate good'.[10] In the modern era, Kant also famously put it that 'autonomy' is a matter of obeying the rational and self-establishing moral principles.

Second, the reasons-responsiveness view fits with our moral practice and ordinary moral intuitions. Related to the first point, the possession of the rational ability seems to only bestow candidacy of responsible agent to those who are capable of rational thinking. We only take normal adults as genuinely responsible and we exclude almost all other animals and children as full-fledged responsible agents. The straightforward explanation for this discrimination is that these other subjects do not possess the mature rational ability as normal adults do. Psychopaths may be a tricky case. Sometimes we do feel reluctant to ascribe moral responsibility to them. However, the reluctance may probably be explained by the fact that we do not know how to assess psychopaths' rational ability appropriately. We do not know, for example, whether they fail to understand the value of moral reasons or they are just indifferent to them. This lends further support the idea that we identify the reasons-responsiveness as a key component of responsible agency. Finally, if we are informed that a person's reasons-responsiveness is (temporarily) impaired, we are very likely to (temporarily) exempt him from being morally responsible.[11]

A related point is that the reasons-responsiveness view fits with our legal practices. Although legal responsibility cannot be equated with moral responsibility, these two concepts are deeply correlated with each other. Thus, studying legal practices provides important insight into how we think about moral responsibility. This is the view held by Nelkin, who is a defender of the rational ability view. Nelkin (2011, chap. 1) reviews the legal data and material about jurors and courts' considerations on imposing the death penalty. She finds out that what typically mitigates the jurors and courts decisions on imposing the death penalty are related to the defendants' mental ability such as having learning disabilities, extreme emotional disturbance and traumatic childhood. She argues that such considerations indicate that when

[10] For a reconstruction of Aristotle's account of responsible agency, see Irwin (1980). Irwin argues that from Aristotle's writing, two different views about responsible agency can be constructed—the simple theory and the complex theory. Irwin thinks the complex theory is more tenable, which can be characterized as "A is responsible for doing X if and only if (a) A is capable of deciding effectively about x, and (b) A does X voluntarily".

[11] Nelkin ( 2005; 2011, chap. 1) provides some interesting examples for this point. Many philosophers worry that the situationist research in social psychology (e.g., the Stanford Prison Experiment and Milgram's Experiment) pose a challenge to our ordinary ascription of moral responsibility. For this research shows that our rational ability may be impaired by unexpected situational factors.

jurors make decisions about legal punishment, they care much about the defendants' rational ability. Though Nelkin admits that 'the relationship between responsibility and punishment is far from clear' (2011, 12), she holds that our legal practices support the reasons-responsiveness view.

There should be no surprise that responsibility and reasons-responsiveness bear a tight relation if we consider the nature of holding others responsible. According to an influential account developed by Peter Strawson (1962), to view a person as a responsible agent is to take her as an appropriate target of our reactive attitudes. Here reactive attitudes refer to a range of moral emotions such as gratitude, resentment and guilt. Strawson's account brings us to a further question—what it is meant by taking one as an appropriate target of our reactive attitudes. A plausible suggestion is that to take an agent to be the appropriate target of reactive attitude is to hold the agent to a specific kind of expectations—the expectations that i) she will comply with the moral obligations accepted by members of the moral community, and that ii) she can understand and communicate with the reactive attitudes conceptually associated with the moral obligations.[12] An agent can meet the expectations of i) and ii) only if the agent possesses the ability to understand certain moral reasons and govern herself with those moral reasons. That is, by looking at the nature of holding responsible, we can explain why reasons-responsiveness entitles a person to be a responsible agent in our moral community.

And finally, being a compatibilist theory, the reasons-responsiveness theory preserves one of our important intuitions—our moral practice seems not to hinge on the metaphysical-empirical question whether our universe is deterministic.[13] Fischer vividly makes this point:

> I could certainly imagine waking up some morning to the newspaper headline, "Causal Determinism Is True!" … And I feel confident that this would not, nor should it, change my view of myself and others as (sometimes) free and robustly morally responsible agents—deeply different from other animals. The mere fact that these generalizations or conditionals have 100 percent probabilities associated with them, rather than 99.9 percent (say), would not and should not have any effect on my views about the existence of freedom and moral responsibility. My basic views of myself and others as free and responsible are and should be resilient with respect to such a discovery about the arcane and "close" facts pertaining to the generalizations of physics. (Fischer, 2007, 44-45)

[12] For this point, see Watson 2012; Wallace 1996; McKenna 2013.
[13] Here I leave it as an open question whether determinism is a matter of empirical sciences or metaphysics. It will not affect the point either we take it as a metaphysical one or an empirical one.

Fischer's point is that our attitudes towards moral responsibility should be *resilient* to the discoveries in sciences and metaphysics. Peter Strawson (1962) has raised a similar point. In his seminal paper 'Freedom and Resentment', he argues that our moral practice is immune to metaphysical import such that metaphysical convictions by no mean can change our basic moral responsibility practices such as blaming and praising. If our ordinary moral practices and attitudes are not sensitive to the endorsement of empirical results and metaphysical positions, then our responsible agency should also be grounded in a more mundane ability that is compatible with different empirical or metaphysical theses.

In summary, the general reasons-responsiveness theory is a defensible and promising proposal to specify how free and responsible agency exists even if the thesis of determinism is true.

### 2.2. The Problems with Reasons-Responsiveness

#### 2.1.1 The Incompatibilist Challenges?

Many incompatibilists doubt that reasons-responsiveness can capture the idea of free agency. They hold that genuine free agency requires certain demanding metaphysical conditions to be satisfied (e.g., we have categorical ability to do otherwise, or we have ultimate control over our characters, or we can agent-cause our actions). These metaphysical conditions certainly go beyond the requirement of reasons-responsiveness.[14] For the purpose of the thesis, I am going to set aside most of these incompatibilisits objections and assume that the reasons-responsiveness theory is adequate to capture the free agency which is required by moral responsibility. I think this omission of those incompatibilist objections is justified.

First, there have already been many exchanges between incompatibilists and compatibilists. Reasons-responsiveness theorists have sophisticated replies to these challenges. In general, they think that those demanding conditions postulated by incompatibilists are not relevant to the free agency that we want (say, the free agency which grounds moral responsibility).[15] To do justice to these debates would exceed the scope of the thesis.

[14] This problem is particularly raised from the incompatibilist camp. Specifically, many incompatibilists worry that reasons-responsiveness is not sufficient for moral responsibility because to have free will and moral responsibility, one has to have ultimate control over his action. It is difficult to see how the reasons-responsiveness theory can accommodate this condition. For this concern, see Kane (1996) and Pereboom (2001).

[15] For example, reasons-responsiveness theorists deny that ultimate control is required by moral responsibility. E.g., Fischer 2012, chap. 10.

Second, the disagreement between compatibilists and incompatibilists is probably not just a metaphysical disagreement and cannot be settled merely in metaphysics. Since there are different definitions of free will, philosophers usually use moral responsibility as the anchor point for the debate: that is, an agent enjoys genuine freedom only if it is justified to ascribe moral responsibility to her. Therefore, the disagreements about freedom often amount to a disagreement on the conditions for moral responsibility. However, the thought that moral responsibility (and the related intuitions) serves as a litmus paper for genuine free agency usually hinges on another supposition. That is, there is a simple and unified conception of moral responsibility. This presupposition is doubtful. The conceptual geography of moral responsibility may be more complicated than previously expected. For example, the notion of desert occupies a central role in understanding moral responsibility. A person is morally responsible for what he does when he deserves the praise, blames or punishment resulting from his action. But what is desert? Compatibilists and incompatibilists may have different answers in mind. To say one is deserving the anger from his best friend for the wrongful action, for example, is quite different to say that he is deserving certain kind of punishment. Thus, I submit that the disagreement between the compatibilists and incompatibilists must lie outside the realm of metaphysics and that it cannot be settled without a deeper understanding of the practice of moral responsibility, say whether blame is exclusively retributivist, and whether a blameworthy action entails justified punishment. The discussion of the nature of moral responsibility, of course, will go beyond the scope of the thesis.

### 2.1.2 Reasons-Responsiveness and the Frankfurt-Style Cases

I think the most pressing problem with reasons-responsiveness ironically comes from the compatibilist side. That is the Frankfurt-Style cases—a kind of thought experiments which were first introduced by Harry Frankfurt (1969). The very aim of these thought experiments is to help compatibilists to deal with the challenge from determinism. However, the argument from the Frankfurt-Style cases is a double-edged-sword—it equally causes trouble to compatibilism, particularly, the reasons-responsiveness theory.[16]

The basic conclusion that the Frankfurt-Style cases try to establish is that moral responsibility does not require the freedom to do otherwise. Since the freedom to do otherwise is what determinism excludes, this

[16] I partly agree with Vihvelin's judgement that "if [Frankfurt's] aim [to introduce the Frankfurt-Style cases] was to make it *easier* to defend compatibilism, he has failed." (Vihvelin 2000, 2, my emphasis) However, as I will show in chapter 2, the Frankfurt-Style cases do provide the compatibilists with a dialectical advantage, though in a circuitous way.

inspires a generation of compatibilists who think that the most important and valuable free will is compatible with determinism. Here is a typical Frankfurt case:

> Jones is an agent confronting a hard choice whether he will cheat on the exam. Unbeknownst to Jones, a resourceful neuroscientist, Black wants to ensure Jones cheats on the exam. Black has implanted a tiny chip into Jones brain when Jones was asleep. This chip can not only monitor Jones's brain activities but also make Jones decide in accordance with Black's will. Black is going to make Jones decide to cheat only if he discovers that Jones is inclined not to cheat. Otherwise, Black will let Jones make his own decision. It turns out that Jones decides to cheat on his own.

In the above scenario, Jones has no freedom to do otherwise rather than cheating because of the presence of Black. Still, we have a strong inclination to attribute moral responsibility to Jones.[17] Call it the Frankfurt-Style cases intuition. This intuition speaks against the well-trenched principle in the discussion of free will and moral responsibility, which is now known as the principle of alternative possibilities (PAP):

> **PAP**: a person is morally responsible for what she has done only if she could have done otherwise.[18]

The Frankfurt-Style cases provide compatibilists with a significant dialectical advantage in the debate with incompatibilists. Recall that incompatibilists hold that determinism is threatening to our free will because it rules out our ability to do otherwise. If it turns out that the ability to do otherwise is not necessary for the freedom required by moral responsibility, then the worry about determinism will be eased to a great extent.

However, the Frankfurt-Style cases also pose challenges to the classical compatibilists proposals which identify freedom with the (conditional) ability to do otherwise. According to these proposals, the agent is able to do otherwise if certain counterfactuals are true (say, the agent would do otherwise if he desired or chose to do otherwise). Such conditional ability is ruled out by the counterfactual intervener (the neuroscientist) in the Frankfurt-Style cases. Likewise, the Frankfurt-Style cases cause trouble to the reasons-responsiveness theory. For according to a straightforward conception, namely, the leeway approach, reasons-responsiveness also consists of the conditional ability to do otherwise. The Frankfurt-

[17] There is a huge debate about whether the Frankfurt-style cases succeed to establish the purported conclusion. Since Frankfurt's original paper, many philosophers have devised more sophisticated versions of Frankfurt-Style Cases. For a collection of the discussion, see Widerker and McKenna (2003).
[18] See Frankfurt (1969, 829). Frankfurt's original term for this principle is 'the principle of alternate possibilities'.

Style cases then indicates that reasons-responsiveness is not necessary for moral responsibility, which speaks against the leeway approach to reasons-responsiveness. I call it the *Challenge of Unnecessity*.

Besides, the reasons-responsiveness theory confronts another challenge from the Frankfurt-Style cases. More specifically, the reasons-responsiveness theory is in tension with an influential model of free agency inspired by the Frankfurt-Style cases. As mentioned, the Frankfurt-Style cases trigger the intuition that the agent in the scenario is morally responsible despite lacking access to alternative possibilities. Many compatibilists further explain this intuition by appealing to a more general principle, namely, the actual-sequence view of freedom. According to the actual-sequence view, whether an agent is acting freely depends on the actual-sequence that issues in the action. Specifically, if the actual event sequence is the right one (say, the event sequence leading to the action involves no responsibility-undermining factors such as mental disorders, manipulation, hypnosis, coercion and so on), then the action is free; and vice versa, the action is not free. With the actual-sequence view of freedom, we only need to look at the actual causal determinants of an action to ascertain whether the agent did it of their own free will, and in a way that can ground assessments of moral responsibility.

When the actual-sequence view is taken more seriously, it turns out to be in tension with the reasons-responsiveness theory. Recently, Sartorio (2015; 2016) proposes that this actual-sequence view should be read as a grounding claim about freedom: that is, free agency should be *exclusively* grounded in the actual causal history; thus, anything which does not pertain to the actual causal history does not help to ground freedom. Sartorio further argues that under this interpretation, the actual-sequence view cannot square with the reasons-responsiveness theory. Specifically, reasons-responsiveness is a model property which is specified in terms of counterfactual possibilities; it is difficult to see how it can be reflected in the actual causal history of the action. It should be noted that the tension between reasons-responsiveness and the actual-sequence view is different from the tension between reasons-responsiveness and the Frankfurt-Style cases intuition. The Frankfurt-Style cases intuition implies that reasons-responsiveness (and the ability to do otherwise) in unnecessary for freedom and moral responsibility; while the actual-sequence view implies that reasons-responsiveness is irrelevant to freedom and moral responsibility because reasons-responsiveness does not play any roles in grounding freedom and moral responsibility. Thus, I call the latter the *Challenge of Irrelevance.*

Due to these two challenges from the Frankfurt-Style cases, the compatibilists split into two—the leeway compatibilists and the source compatibilists. The leeway compatibilists deny the Frankfurt-Style case intuition and they inherit the classical compatibilist proposal which identifies freedom as the agent's ability specified in conditionals; by comparison, the source compatibilists accept the Frankfurt-Style case intuition and the actual-sequence view according to which freedom is the right causal history leading to

the action. That is to say, there are currently two different compatibilists strategies which are potentially in tension with one another. One of the goals of the thesis is to help the reasons-responsiveness theory with these two challenges posed from the Frankfurt-Style cases (I will deal with these two challenges in Chapter 2 and Chapter 3 respectively). By addressing these two challenges, the thesis helps to show that the leeway compatibilist proposal and source compatibilist proposal can be shown to be compatible and unified. Besides, as to be shown below, to the project of integrating a reasons-responsiveness theory and the causal theory of action hinges on the success of tackling of the Challenge of Irrelevance.

### 2.1.3 The Absence of an Action Theory

Traditionally, the reasons-responsiveness theorists focus on the psychological aspect of free agency—being free is being rational and mentally healthy; while little attention has been paid to the notion of action. Still, the ability to respond to reasons hinges on the very ability to act. Without an account of intentional action, the picture of free agency is incomplete. Given that the notion occupies a central role in an account of reasons-responsiveness, it is surprising to see that reasons-responsiveness theorists seldom discuss the issues surrounding action theory. I think part of the explanation is that they want their theories to be non-committal. For example, Fischer introduces his theory of reasons-responsiveness (or "guidance control") in the following ways:

> I contend that my accounts of guidance control (and moral responsibility) are compatible with a wide range of plausible views about these contentious empirical and philosophical matters. For example, my account of guidance control certainly does not presuppose that there is irreducible, indeterministic agent-causation; it thus does not depend, for its acceptance, on some sort of defense of this highly contentious doctrine. On the other hand, I believe that the core of the account is compatible with the existence of irreducible, indeterministic agent-causation. As with Ginet's suggestion, there would perhaps need to be certain adjustments or clarifications; but there is nothing in the core ideas of the account that requires either the truth or falsity of claims about agent-causation. (Fischer 2012, 15–16)[19]

Fischer deliberately leaves many of the details unfilled. The aim is to make his theory compatible with as many other metaphysical and empirical positions as possible. There are certain advantages to being non-committal. First, it enables the theory to be immune to possible challenges from empirical or

[19] Apart from action theory, Fischer contends that his theory is also neutral on a wide range of issues, such as materialism of mind, reductionism in metaphysics, and the ontological status of reasons. (Fischer 2012, p.15)

metaphysical discoveries; Second, being non-committal makes the theory easy to be integrated with other theories, such as the agent-causation theory of action, as indicated in the quote.

Contra Fischer's view, I think there are stronger reasons to combine the reasons-responsiveness theory with a specific account of action. Firstly, as said, the ability to respond to reasons is based on the ability to act for reasons. To have a complete understanding of free agency, we require an account of action. More importantly, since our main concern is whether free agency exists, we should not be satisfied merely by an account about how free will and determinism are compatible. Recall that the challenge from the scientific picture is twofold—one is from determinism and the other is from the universal event-causal assumption. Suppose that the manifest image of agency commits to an agent-causation account of action which turns out to be incompatible with the scientific image, we then still have no good reasons to believe the existence of free agency. Without showing our ordinary conception of action can be obtained in the scientific image, our defence of free agency is incomplete.

In my thesis, I am going to argue that to tackle the challenges from the scientific image, the reasons-responsive theory should be united with an account of action, specifically, the causal theory of action. This account is popular because it promises that action is constituted with an event-causal process, which is compatible with the scientific image. Actually, many proponents of reasons-responsiveness theory have already explicitly or implicitly endorsed the causal theory of action. For example, Michael McKenna, another defender of reasons-responsiveness theory, expresses his commitment to the causal theory of action explicitly.[20] And Markus Schlosser defends the causal theory of action and the reasons-responsiveness theory respectively (Schlosser 2010, 2013). Even Fischer and Ravizza, who deliberately stay non-committal on the issues related to action theory, seem to take the causal theory of action as a default position for an account of action.[21]

Proponents of reasons-responsiveness, however, do not bother specifying how the reasons-responsiveness theory and the causal theory of action can fit together. Why they do not take more time to justify their stances on action theory? One explanation is that they think an action theory, in particular, the causal theory of action, will come easily to fill the gap left by a reasons-responsiveness theory. As Fischer suggests in the text quoted above, his theory can be integrated with the agent-causation theory of action with just minor clarifications and adjustments. I think many of the defenders of reasons-responsiveness

[20] He writes that "given my allegiance to naturalism, I endorse a causal theory of action" (McKenna 2012, 18).
[21] For example, in a passage clarifying the relationship among reasons-responsiveness and intentional action, they seem to presume that intentional actions are more basic agentive abilities than reasons-responsiveness. (That is to say, if an agent is acting in a reasons-responsive way, then the agent must be acting intentionally.) And when they characterize intentional action, they write that "an action is intentional in the sense that it is produced (in an appropriate way) by the [agent's] beliefs and desires" (Fischer and Ravizza 1998, 82).

theory will have similar opinions—if, with certain clarifications and adjustments, the reasons-responsiveness theory can be shown to be compatible with the agent-causation account of action, then with equal or even fewer clarifications and adjustments, the reasons-responsiveness theory can be shown to be compatible with the causal theory of action. For the causal theory of action commits a simpler and more intelligible causal order than the agent-causation theory.

However, contra this common expectation, a causal theory of action will not be integrated with the reasons-responsiveness theory for free. Recall that as Sartorio suggests, the reasons-responsiveness theory is in tension with the actual-sequence view: it is difficult to see, how, reasons-responsiveness, as a modal property, can be reflected in the actual causal history. This difficulty also applies to the integration project because the causal theory of action and the actual-sequence view has a similar structure. Suppose we combine the reasons-responsiveness theory and the causal theory of action—acting freely, can be cashed out in terms of causal interaction of events. According to this combined theory, a free and responsible action is an action performed in a reasons-responsive way, which can be cashed out in terms of the actual causal history. This actual causal history is probably constituted with agent's deliberational processes and his mental states such as beliefs and desires. Now we can ask the same question—how is reasons-responsiveness, which is understood as a modal property, reflected in the actual causal history? Thus, to pave the way for the integration project, what we first need to do is to show that reasons-responsiveness, even though understood as a modal property, is reflected on the actual causal history of the action. And this amounts to tackling the challenge of Irrelevance posed by the Frankfurt-Style cases and the actual-sequence view.

## 3. Integrating the Causal Theory of Action

### 3.1 What is a Causal Theory of Action: The Aim of the Theory

I now turn to an overview of the causal theory of action. The notion of 'causal theory of action' actually covers a bunch of theories with variation of details (Davidson 1963; Goldman 1970; Brand 1984; Bratman 1987; Bishop 1989; Mele 1992). This theory is about the nature of action. Broadly speaking, the causal theory of action (henceforth CTA) is the view that actions are bodily movements caused by our motivating mental states such as desires, beliefs, intentions. In particular, these motivating mental states can serve as a rational explanation for the action. I call this the standard version of CTA. There are three contentions which make the standard CTA differ from alternative accounts of action in the literature. First of all, according to the standard CTA, actions by nature are causal phenomena. With this feature, the standard CTA differs from the non-causal action theories (e.g., Goetz 1988; Ginet 1990) according to

which action (or basic action) has no causal structure or cannot be understood as causal phenomena. Second, action is to be analysed into event-causation. This feature distinguishes the theory from agent-causation accounts (e.g., Chisholm 1966; O'Connor 2000), according to which action is caused by the agent as an unreducible substance. Third, action is not only analyzed into events, but also into non-actional events. This feature distinguishes the standard CTA with theories which invoke actional events such as volition or trying. (e.g, Hornsby 1980; Ginet 1990)

The standard version of CTA, of course, can be read in an intensional way and an extensional way, depending on what kind of aim we want to achieve through the theory. From the intensional reading, CTA is providing a conceptual analysis for action. For example, the analysans 'justified true belief' is a proposal to analyse the concept of knowledge (though it is commonly thought to be nonsufficient due to the Gettier Problem). Likewise, the analysans "bodily movement being caused by specific mental states in a certain way" can be taken as a proposal to analyse the concept of action. Thus, to see whether CTA is an adequate analysis of the action, we should investigate how ordinary people use the concept of action.

By comparison, from the extensional reading, CTA aims to provide a co-extensional judgement between action and the causal process at issue. Simply put, action is what happens when the concept of action has application. This reading is popular in treating concepts of natural kind terms. For example, 'water is $H_20$' is not an illustration of how the concept of water is used by ordinary people; Rather, it is a claim about molecular nature of the watery stuff which we call water. Likewise, the causal theory of action is not providing an explication of the concept. Rather, it is providing the underlying process of when people act.

Which reading of CTA should be endorsed? The defenders of CTA seldom make this point clear.[22] I think neither the intensional nor the extensional reading is satisfactory. The complication of the issue is that 'action' is not like a term that is defined as human interest, nor it is like a natural kind term. First of all, we care more about what action is rather than what action means. More specifically, we care about the metaphysical issues with action—that is, whether action as we ordinary understands it finds its place in the scientific image. Therefore, I do not think that the intensional reading is adequate.[23]

The extensional understanding is problematic also. For the extension of the concept of action can be settled merely through empirical investigation, regardless of the ordinary conception of action. If the CTA fails to accommodate our ordinary concerns about action, then it does not contribute anything to the

[22] Michael Moore (2010) and Mayr (2011) have similar remarks on the ambiguity of the CTA program.
[23] Some philosophers argue against CTA by defending the claim that action and intentional action is conceptually primitive. (e.g., Ford 2011; O'Brien 2017) However, if we do not take CTA as an enterprise of conceptual analysis, then many of their criticisms seem to misfire.

reconciliatory project, namely, reconciling the scientific and manifest images of agency. More specifically, the problem can be spelt out from two respects, one is explanatory, and the other is ontological.[24]

Imagine that through empirical investigation we learn that human action always involves a certain kind of psychological process, which is further underpinned by a certain kind neural-biological mechanism. But still, we might not understand how agentive control comes about from the causal process in question. That is to say, there is an explanatory gap between these neural-biological mechanisms and the ordinary conception of action.[25] This is the *explanatory problem* with the extensional reading of CTA. Besides, even if we find out that the occurrence of action always involves a specific causal process, it does not follow that action is nothing more than that causal process. It is possible that action is just accompanied by the causal process, while the nature of action (say, agent-causation) is still beyond the scientific image.[26] This is the *ontological problem* with the extensional reading of CTA.

These two problems correspond to two requirements of a satisfactory CTA analysis: first, a satisfactory CTA analysis needs to facilitate our understanding about how action can come about through the event-causal process. And second, it needs to show that action is *nothing over and above* the causal process provided in the analysis. I propose that the aim of CTA is to provide a reductive explanation for action. Or more specifically, to provide a reductive explanation for the phenomena of acting (as it is presented in the manifest image) in terms of causal interactions among events.[27] Thus, I characterize CTA in the following way.

> CTA as a reductive explanation: an action obtains in virtue of a specific event-causal process which is constituted with mental states and bodily movement in a specific structure.

[24] This is parallel to the problems with the 'Nagellian reduction' for mental properties. See Kim 1997, chap. 4.
[25] There is a similar problem with physicalism in philosophy of mind. Just pointing out that there are certain physical processes accompany the mental processes is inadequate to explain how the mental properties (say, intentional properties and qualitative properties) come about from the physical properties.
[26] Bishop (1989, chap. 3) helpfully makes a distinction between an underlying process contingently correlating with action and the process constituting the action. It is the latter that a proponent of CTA should argue for. More on this point below. See also Mayr 2011, chap. 5.
[27] This idea of reductive explanation is borrowed from the discussion in philosophy of mind. It is a popular idea to take physicalism as a reductive explanation—to reductively explain mental properties (intentional and phenomenal) in terms of physical properties. See Chalmers 1996, 42–47; Kim 1998, 97–103.

From this understanding, the key commitment of CTA is a relation of ontological dependence between action and the relevant causal structure. Particularly, this relation must be able to explain why action—as we ordinarily understand it—can arise through a specific event-causal process.[28]

But how is this reductive explanation of action advanced? We should first look at a similar proposal by Bishop. Bishop agrees that CTA should be understood and defended as a claim that action ontologically depends on the event-causal process. Specifically, action is *realized* and *constituted* by the event-causal interactions between mental states and bodily movement.[29] According to Bishop's suggestion, to defend CTA is to show that the causal process serves as a set of sufficient and necessary conditions for action. Importantly, Bishop adds that this set of conditions not only apply to our actual world, but to all other possible worlds which are similar to the actual world in relevant respects (say, with the same causal orders and ontology).[30] He hopes that by making such a modal stipulation, the CTA analysis can guarantee that action is *constituted* rather than just *correlated* with the causal process.[31] However, this modal stipulation seems not to fulfil Bishop's expectation. Even if we have a set of sufficient and necessary conditions for action which applies to all possible words with similar respects, it does not follow that action is constituted and realized by those conditions. Here is a counterexample: an object's having colour is sufficient and necessary for the object's having shape in all relevant possible worlds. However, it does not follow that having shape is constituted by having colour.

Contra Bishop's suggestion, the crucial point of reductive explanation, is not to look for necessary and sufficient conditions. Rather, it is to provide a functional analysis of the ordinary concept of action.[32] A functional analysis of a phenomenon or property is to specify its functional roles within a system or framework, where its functional roles are understood as the relations to other properties or phenomena, typically being cashed out in causal or dispositional terms. Functional analysis serves as the starting point for reductive explanation. Only with a proper functional analysis can we see whether the target phenomenon or property can be realized and reductively explained by more basic phenomena or properties. Here is an example. To reductively explain what gene is, we first need to provide a functional analysis of gene—that is, roughly, something which enables organisms to pass on biological traits to their

[28] There are several candidates for the relation of ontological dependence, such as realization, grounding, and logical supervenience. I take it as an open question which candidate should be endorsed as an interpretation for this relation.
[29] See Bishop 1989, 96. Following Bishop, Clarke also use 'realization' to characterize the ontological dependence. Clarke further construes realization as an 'be an asymmetric (and hence irreflexive) relation of ontological dependence and determination' which is thought to be allowed multiple realization (Clarke 2019, 753).
[30] Bishop 1989, 97.
[31] The aim of this move is to tackle the ontological problem. See note 26.
[32] Bishop sometimes suggests that the concept of action should be understood as a functional concept (Bishop 1989, 143). In addition, his account is influenced by a functional conception of control from cybernetics. However, Bishop never seriously take up the enterprise to provide a functional analysis for action (Bishop 1989, 168).

offspring. And through empirical research, it turns out that this functional role is realized by specific DNA sequence, the double helix. Then we get a reductive explanation for gene. Likewise, I submit that to reductively explain action in terms of an event-causal process, what we first need to do is to construe the concept of action in a functional way. In particular, as I will show, the essential part of the functional analysis of action is the functional analysis of some essential characteristics of action, namely, Control-Maintenance and Agent-Participating. As I will show in the following section, the main challenges to CTA can be viewed as challenges to the project of reductively explaining these characteristics of action.

The advantage of the reductive explanation reading preserves some of the features of the intensional reading and the extensional reading. For reductive explanation is a project of both conceptual and ontological investigations. The reductive explanation of action does not require a rigid conceptual analysis of action. Still, it requires a functional interpretation of action. In this part, our common-sense understanding and pre-theoretical intuitions about action should be taken into account. Besides, the reductive explanation does not stop at the common-sense level. It should go further to show that the functional analysis is actually realized by an event-causal process—which amounts to be an ontological investigation.

Apart from understanding CTA as a kind of reductive explanation, there are other advantages to my characterization of CTA. Unlike the standard version of CTA, my characterization does not commit to the idea that actions are bodily movements. Consequently, it sets aside the issues of the identification of action—whether action is identical to bodily movement, or action is identical to the abstract causing between mental and physical, or action is identical to the whole causal process.[33] The second advantage is that this characterization does not indicate that the causal relation between the mental states and the bodily movements (or the action) is a linear process. Rather, it allows a more complicated structure for the event-causal process—which can be hierarchal and on-going. As I will show in the following chapters, this move is crucial to handle certain criticisms to CTA.

### 3.4. The Problems With CTA

CTA aims to provide a solution to the apparent tension between the manifest image of intentional action and the scientific image. Despite being promising to fit with the scientific image, this theory confronts many criticisms. It would not be exaggerating to say that CTA, particularly the standard CTA, is currently the most unpopular theory in the action theory literature. Compare this to the situation of the traditional

[33] Identifying action with the bodily movement will trigger some problems for the CTA. (Haddock 2005; Ford 2011) Hopefully my characterization of CTA can help to avoid such criticisms.

definition of knowledge: everybody knows that the justified-true-belief analysis of knowledge is problematic. Still, to those who do not work on the particular issue concerning the definition of knowledge, this analysis serves as a working hypothesis for what knowledge is. The standard CTA is in a similar situation. Almost everybody knows that it is problematic—action cannot be simply understood as a bodily movement caused by motivating mental states. Still, it becomes a rule of thumb to help philosophers working in areas other than action theory when they need to sketch what action is.

By why is CTA so unpopular in action theory? I have shown that CTA can be viewed as a reductive explanation for action in terms of event-causation. However, there are certain characteristics of action in the manifest image which resist such a reductive explanation: an agent is maintaining control during an action; and that an agent is participating and fulfilling certain roles in her action. These two characteristics correspond two problems that plague the causal theory of action—the problem of causal deviance, and a particular version of the problem of disappearing agency.

**The Problem of Causal Deviance**

According to the standard CTA, a bodily movement can be counted as an action if and only if it is caused by specific mental states in a specific structure. However, there are certain counter-examples which now known as the cases of causal deviance. Consider the famous example composed by Donald Davidson:

> A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never chose to loosen his hold, nor did he do it intentionally. (Davidson 1973/2001, 79)

In this case, the climber's loosening the hand is caused by his mental events (belief and desire). However, this motion from the climber does not count as his action.

To handle the cases of deviance, defenders of CTA try to provide more sophisticated analysis for action. Several suggestions have been proposed (e.g., Peacocke 1979; Bishop 1989). However, there are always new deviant cases contrived to bypass the extant proposals. Based on this status quo, opponents of CTA suggest that such a failure is an indicator that CTA cannot fully capture the concept of action. I will tackle this problem in Chapter 4.

**The Problem of Disappearing Agency**

Another difficult problem confronting CTA is the problem of disappearing agency, according to which CTA makes the agent disappear from the occurrence of action. Although several opponents of CTA have raised the problems of disappearing agency in various ways, the most classic characterization is from David Velleman's paper 'what happens when someone acts' published in 1992:

> I think that the standard story is flawed in several respects. The flaw … is that the story fails to include an agent-or, more precisely, fails to cast the agent in his proper role. In this story, reasons cause an intention, and an intention causes bodily movements, but nobody-that is, no person-does anything. Psychological and physiological events take place inside a person, but the person serves merely as the arena for these events: he takes no active part. (Velleman 1992, 461)

Even though this problem is repeatedly raised as a criticism to CTA, philosophers disagree on what this problem means and how serious a challenge it poses to CTA. I am going to address the problem of disappearing agency in chapter 5 and Chapter 6 of the thesis.

Both the problem of causal deviance and the problem of disappearing agency are rooted in the scientific image discussed in the former sections, according to which all causal interactions are fundamentally event-causal interactions and human agency is no exception to the phenomena of the natural world. Similar to the determinist challenge to free will, the problem of causal deviance and the problem of disappearing agency are on frontline of relieving the tension between the scientific image and the manifest image about human agency. In my thesis, I contend that both problems are not insuperable and that with the resources from the reasons-responsiveness theories, the causal theory can better handle the problems.

**4. The Agenda of the Thesis**

In Chapter 2, I will establish a general conception of reasons-responsiveness. That is, reasons-responsiveness consists of the agent's ability to do otherwise (conditionally understood). The main obstacle of defending this view is the Frankfurt-Style cases, which triggers the Challenge of Unnecessity (namely, that the ability to do otherwise is not necessary for free agency). Because of this challenge, there are two attempts to dissociate reasons-responsiveness with the agent's ability to do otherwise. The first one is Fischer and Ravizza's mechanism-based reasons-responsiveness account; the second is Sartorio's causal reasons-sensitivity account. I criticize both accounts. Then I provide a proposal to reconcile the

view that reasons-responsiveness consists of the ability to do otherwise and the basic intuitive judgment elicited by the Frankfurt-Style cases.

In Chapter 3, I will address a pressing and recent objection to reasons-responsiveness views, namely, the Challenge of Irrelevance, and show how the actual-sequence view and the modal conception of reasons-responsiveness are compatible. I argue that both modal and non-modal conception is required to understand (different aspects of) human agency. This also paves the way to the combination of the reasons-responsiveness theory and the causal theory of action.

Then I will turn to the causal theory of action and the challenges confronting this theory. In Chapter 4, I will tackle the problem of causal deviance. In doing so, I develop a novel account of control. This account takes into considerations several factors which may influence our judgment whether a certain object is in control. These factors include causation, purposiveness, accuracy, reliability and flexibility. According to this account, our conception of control is multi-faceted and control is a notion comes in degrees. An implication is that it is neither possible nor necessary to draw a sharp line between the deviant causal chains and the non-deviant causal chains. In effect, the problem of the causal deviance is dissolved.

Chapter 5 and Chapter 6 are about the problem of disappearing agency. I separate the discussion of this problem into two chapters because there can be different interpretations and reconstructions for this problem. In Chapter 5, I will first enumerate several possible interpretations for the problem of disappearing agency, and then I will point out certain conceptual issues related to the problem which I think are the most pressing. Specifically, under one interpretation, the problem is that the event-causal framework employed by CTA cannot accommodate two intuitions about action, namely, Agent-Participation (that agent participates in her action) and Anti-Reduction (that agency as active phenomena cannot arise from merely event-causal interactions). To tackle this problem, I will develop a positive account, namely the structural account, to show that agent's roles in her action can be fulfilled by the causing of her mental events as well as her psychological structure. I will also provide arguments to explain away the intuition of Anti-Reduction.

In Chapter 6, I will address the problem of disappearing agency under another interpretation according to which the causal theory of action cannot accommodate certain kinds of phenomenology of agency. I will examine several kinds of phenomenology which are purported to involve content about agent-causation or non-event-causation. These include the phenomenology of acting, the phenomenology of making choices and the phenomenology of exerting effort. I will argue that none of this phenomenology will pose a real challenge to CTA.

In the concluding chapter, I will summarise the arguments in the thesis and extract the basic take-home message. That is, we have an account of free agency, which, on the one hand, does not fail to meet our ordinary expectation delivered by the manifest image; on the other, is compatible with the scientific image. This account maintains that:

> (i.) Free Agency (which is required by moral responsibility) is the ability to respond to reasons.
>
> (ii.) To exercise free agency is to act for reasons; while acting for reasons can be captured within an event-causal framework.

## Chapter 2: Reasons-Responsiveness and The Frankfurt-Style Cases: A Middle Path

**Abstract:** In this chapter, I establish a general conception of reasons-responsiveness. That is, reasons-responsiveness consists of the agent's ability to do otherwise. The main obstacle of defending this view is the Frankfurt-Style cases, which I have reviewed in the introductory chapter. Inspired by these cases, there are two attempts to dissociate reasons-responsiveness with the agent's ability to do otherwise. The first one is Fischer and Ravizza's mechanism-based reasons-responsiveness account; the second is Sartorio's causal reasons-sensitivity account. I criticize both accounts. Then I provide a proposal to reconcile the view that reasons-responsiveness consists of the ability to do otherwise and the basic intuitive judgment elicited by the Frankfurt-style cases.

### 0. Introduction

Human agents are morally responsible for what they do in virtue of a specific kind of free agency. Philosophers have long been intrigued by the question of what the nature of this agency is. There are roughly two different ideas regarding this question— incompatibilism and compatibilism. The incompatibilists contend that our responsible agency, if exists, must be an ability or (a set of abilities) which is not compatible with determinism; while compatibilist thinks that this agency is compatible with determinism. One popular compatibilist proposal is to identifies our responsible agency with our ability to recognize reasons and act in light of reasons.[34] Philosophers use different terms to refer to this rational ability (e.g., reasons-responsiveness, reflective self-control and reasons-sensitivity). In this chapter, I will follow the usage of term in Fischer and Ravizza (1998) and refer to this rational ability as *reasons-responsiveness* because this term is most widely used in the literature.

There are currently two competing approaches for the nature of reasons-responsiveness—the *leeway approach* and the *source approach*. The main aim of this chapter is to defend a compatibilist version of the leeway approach to reasons-responsiveness, according to which reasons-responsiveness consists of the agent's ability to do otherwise. This account is composed of two claims: *(i)* that reasons-responsiveness is an ability attributed to the agent and *(ii)* that the notion of reasons-responsiveness is analyzed in terms of counterfactual scenarios. I refer to the first claim as the *agentive ability conception* of reasons responsiveness and the second as the *modal conception* of reasons-responsiveness. The leeway approach contrasts with the source approach to reasons-responsiveness, according to which, reasons-responsiveness

[34] E.g., Wolf 1990; Wallace 1996; Fischer and Ravizza 1998; Vargas 2013; Nelkin 2011.

is not taken as agent's ability to do otherwise; rather, it is a property of the actual sequence of the action. Proponents of the source approach either refute the agentive ability conception of reasons-responsiveness while retaining the modal conception (Fischer and Ravizza 1998; McKenna 2013); or refute both the agentive ability conception and the modal conception (Sartorio 2015; 2016). At a first glance, the leeway approach provides a more natural understanding of reasons-responsiveness. However, the source approach currently attracts more proponents. To see how this dialectical situation evolves in contexts, I briefly review the development of free will debate in the last several decades.

Once upon a time, the most popular understanding of free agency was the ability to do otherwise; while the most popular compatibilist way to analyze the ability to do otherwise is to analyze it with conditionals, e.g., an agent will do otherwise, if the agent chooses/tries to do otherwise. Following this compatibilist schema, a straightforward suggestion is that reasons-responsiveness should also be understood as an ability to do otherwise and be analyzed conditionally. Roughly, a person is reasons-responsive iff, if there is a sufficient reason for her to do otherwise, she will recognize that reason and do otherwise for that reason.

This compatibilist proposal is challenged by the Frankfurt-Style cases, a kind of thought experiments that try to show that the ability to do otherwise is not necessary for the ascription of moral responsibility. Recall, in these kinds of cases, an agent is intuitively responsible for her choice or action, despite lacking the ability to do otherwise. So, even if there were sufficient reasons to do otherwise, she would not be able to respond to those reasons. In other words, in a Frankfurt-Style scenario, the agent is not able to do otherwise, and *ipso facto* not reasons-responsive while still being morally responsible. Usually, the Frankfurt-Style cases are taken as a kind of compatibilist-friendly thought experiments because they provide powerful resources to handle incompatibilism. Recall the main incompatibilist claim is that determinism is incompatible with free will because it forestalls the ability to do otherwise. If, as the Frankfurt-Style cases show, the ability to do otherwise is not necessary for the freedom (as the control condition for moral responsibility), then this incompatibilist claim seems to misfire.[35] However, it has been widely ignored that the Frankfurt-Style cases equally pose a challenge to the compatibilism, as I now describe.

The Frankfurt-Style cases result in a split within the compatibilists. Since the Frankfurt-Style cases were introduced, there have been two conflicting ideas at compatibilists' table. The first one is that free agency is the conditional ability to do otherwise; the other is that freedom (for moral responsibility) does not

[35] I just grant that the focus of the debate is whether the freedom required by moral responsibility is compatible with determinism.

require the ability to do otherwise. Correspondingly, there are two reactions from the compatibilists. The first is to deny the intuition from the Frankfurt-Style cases and insist that the (conditional) ability to do otherwise is necessary for freedom and moral responsibility. This move is taken by the *leeway compatibilists*. Typical representatives are the new dispositionalists, who hold that abilities are either similar to or conceptually connected to dispositions and suggest that abilities are to be analyzed in a similar way to dispositions.[36] Another response is to reject the idea that freedom consists of the ability to do otherwise. This move is taken by the source compatibilists. Inspired by the Frankfurt-Style case, the source compatibilists subscribe to a new model of freedom, namely, the actual-sequence view according to which freedom is a matter of the actual sequence of the action. This view further motivates the source approach to reasons-responsiveness, that is, to embed the notion of reasons-responsiveness into the actual-sequence framework and to dissociate reasons-responsiveness with the ability to do otherwise.[37] This is the story of how the source approach to reasons-responsiveness arose and become popular. Two notable examples of the source approach are the mechanism-based reasons-responsiveness account developed by Fischer and Ravizza (1998) and the causal reason-sensitivity account developed by Sartorio (2015;2016).[38]

In this chapter, contra the source compatibilists, I will argue that reasons-responsiveness consists of the ability to do otherwise. By arguing for this claim, however, I am not siding with the leeway compatibilists either. Contra the leeway compatibilists, I hold that we should endorse the intuition from the Frankfurt-style cases. I am going the middle way by showing that there can be no substantial tension between the leeway compatibilists and source compatibilists—they are not providing competitive answers to the same question; rather, they are providing different answers to different questions. In addition, I will provide a proposal to reconcile the two conflicting ideas. That is, on the one hand, we have a commonsensical idea that free agency consists of the ability to do otherwise; on the other, as the Frankfurt-Style cases suggest, the ability to do otherwise is not necessary for moral responsibility.

Here is the agenda of this chapter. In the first section, I will show that the reasons-responsiveness is an improved version of the conditional ability to do otherwise and how it is challenged by the Frankfurt-Style cases. In the second section and the third section, I will critically review two attempts to dissociate reasons-responsiveness with the ability to do otherwise—Fischer and Ravizza's mechanism-based

[36] For the new dispositionalist position, see Smith (2003), Fara (2008) and Vihvelin (2004; 2013). More on the new dispostionalism in the following section. In addition, Nelkin (2011) also defends a view that the ability to do otherwise survives the Frankfurt-Style case.
[37] Of course, not every source compatibilist is defending the reasons-responsiveness view. For example, Frankfurt (1971).
[38] See also McKenna (2013).

reasons-responsiveness account and Sartorio's *causal reasons-sensitivity* account. Both of these attempts face serious problems. Fischer and Ravizza's mechanism-based approach has the danger of misidentifying the locus for moral responsibility. And Sartorio's account, which tries to de-modalize free agency, fails to capture some essential features of free agency. In the final section, I will provide a way out showing that we can endorse the idea that reasons-responsiveness consists of the ability to do otherwise, while we can accept the direct intuitive judgment from the Frankfurt-Style cases.

## 1. Reasons-Responsiveness as Conditional Freedom to Do Otherwise

### 1.1 The Ability to Do Otherwise and Classical Compatibilism

Traditionally, in the free will literature, the freedom in question is the ability to do otherwise. Roughly, to say an agent is free is just to say he is able to do otherwise than what he actually does. Freedom as the ability to do otherwise is a straightforward idea. Suppose I am sitting at my desk and writing my manuscript. If through these activities I am exercising my free will, it follows naturally that at the same time I am able to choose to do something different, say playing video games or hanging out with my friends. Sometimes, philosophers refer the ability to do otherwise as access to alternative possibilities. In this chapter, I will use the notions *ability to do otherwise, leeway freedom* and *access to alternative possibilities* interchangeably. By alternative possibilities, I mean possible scenarios of agent's performing action that satisfy the two following conditions: i) the possible scenarios are similar to the actual one;[39] ii) in the possible scenario, the agent in question performs a different action than the actual action. Now we can see how the notion of ability to do otherwise and alternative possibilities correlate: if an agent is able to do otherwise, it just means that there are some *similar* possible scenarios in which the agent does something different from what he does in the actual scenario.

This conception of free agency has led many philosophers to find a tension between free will and determinism: If determinism is true, the propositions describing the past and the laws of nature will entail propositions of any events throughout the history, as well as events will occur in the future, which cover whatever the agents actually do. Thus, given the truth of determinism, it seems that we cannot do

[39] By similarity, I mean that both the basic facts (e.g., laws of the nature, the constitution of the world) and the relevant facts (e.g., the agent's psychological and physiological conditions) are similar. However, which facts should be considered highly depends on contexts. As I will show, different ways to understand 'similar' will lead to different positions in the free will debate, notably the dichotomy between compatibilism and incompatibilism.

anything other than what we actually do. This intuitive reasoning motivates one of the most powerful arguments for incompatibilism, namely, the consequence argument.[40]

Confronting the challenge from determinism, classical compatibilists employ a conditional analysis for the ability to do otherwise (e.g., (e.g., Hume 1748/2008; Ayer 1954; Davidson 1973). For example, one can provide such an analysis for the ability to do otherwise in the following way:

> The agent can do X rather than Y, iff if he chooses to do X, she will do X.

The characterization provided above is an example of *simple conditional analysis* for the ability to do otherwise—the agent's doing otherwise is conditioned on the agent's having a different mental antecedent. There are variations of the conditional analyses with respect to the mental antecedents. In such kind of analysis, the mental antecedents can be mental acts, such as choosing, deciding, or trying, or it can be mental states, such as desiring or intending.

It is not difficult to see how the conditional analysis helps to block the consequence argument. The ability to do otherwise, simply put, is the possibility of performing a different action with certain states of affairs held fixed. Different kinds of ability to do otherwise require different parameters being held fixed in the modal analysis. There is no doubt that if determinism is true, then a certain kind of ability to do otherwise is ruled out, namely, the possibility of performing a different action with all the states of the past and the laws held fixed. Call this the *categorical ability to do otherwise*. By comparison, what the classic compatibilist propose is a weaker kind of ability to do otherwise—the possibility of performing a different action when certain mental antecedents of the agent are different. Call this the *conditional ability to do otherwise*. Note that what the conditional ability means is that if the past had been slightly different (specifically, if the agent had different mental antecedents), then the agent would act differently. This is compatible with determinism.

This conditional analysis to some extent captures our intuitive understanding of freedom. Intuitively, one *is* free if one can do what she wants to do; and one is *not* free if he is physically constrained, say being bounded by a rope. Both the positive aspect and the negative aspect can be accommodated by the simple conditional analysis. However, the simple conditional analysis faces counter-examples. Here is one introduced by Keith Lehrer:

> Suppose that I am offered a bowl of candy and in the bowl are small round red sugar balls. I do not choose to take one of the red sugar balls because I have a pathological aversion to such candy. (Perhaps they remind me of drops of blood and …) It is logically

[40] The most detailed presentation of the argument is in Peter van Inwagen (1983)

> consistent to suppose that if I had chosen to take the red sugar ball, I would have taken one, but, not so choosing, I am utterly unable to touch one. (Lehrer 1968, 32)

According to the simple conditional analysis, the agent is free in picking up the candy—for if she chooses or desires to pick up the red candy, she would do so. Intuitively, the agent is not free. She has no opportunities to form relevant desires or intentions because her psychological faculty does not function in the right way. This reveals the shortcoming of classical compatibilism. An agent's freedom will not only be undermined by external and physical impediments, but it can also be undermined by factors that come from the 'inside' – such as pathological aversions, phobias etc. It is these internal freedom-undermining factors that cannot be excluded by the simple conditional analysis.

### 1.2 Reasons-Responsiveness as A Qualified Ability to Do Otherwise

One possible move to improve the classical compatibilism proposal is by making some qualifications to the simple conditional analysis: an agent can do otherwise just means she would do otherwise if she chooses/wants/intends to do otherwise and that she does not suffer from psychological problems X, Y, Z, etc. This list, can, of course, be extremely long. A more elegant proposal is to provide positive conditions for human psychology to function properly rather than excluding all the psychological problems which might undermine freedom. Here is how the notion of reasons-responsiveness comes into play. It is a natural suggestion that by well-functioning psychology, we just mean 'to be reasons-responsive'. To circumvent the counterexamples of psychological disorders or psychological incapacities, compatibilists require a hypothetical analysis of reasons-responsiveness. Very roughly, if an agent is reasons-responsive in doing A, then in possible scenarios in which everything is similar but there are sufficient reasons not to do A, the agent would respond to the reasons and refrain from doing A. According to this approach, reasons-responsiveness is understood as agent's ability to do otherwise and it is analyzed conditionally. I call it the leeway approach to reasons-responsiveness. Unlike classical compatibilism which takes the ability to do otherwise as the constrained possibility of performing a different action when there is a different mental antecedent, this leeway approach to reasons-responsiveness takes the ability to do otherwise as the constrained possibility of performing a different action when there is a different sufficient reason.

### 1.3. The Tension Between the Frankfurt-Style Cases and the Leeway approach

Both classical compatibilism and the leeway approach to reasons-responsiveness are built on the idea that freedom consists of the agent's conditional ability to do otherwise. This idea, however, confronts a challenge. In 1969, Harry Frankfurt devised a seminal kind of thought experiments, which are now known as the Frankfurt-Style cases. From those thought experiments, Frankfurt tries to conclude that moral responsibility does not require the freedom to do otherwise. Recall that we encountered a typical Frankfurt-Style case in the previous chapter. Let's review the case here:

> Jones is an agent confronting a hard choice whether he will cheat on the exam. Unbeknownst to Jones, a resourceful neuroscientist, Black wants to ensure that Jones cheats on the exam. Black has implanted a tiny chip into Jones brain when Jones was asleep. This chip can not only enable Black to monitor Jones's brain activities but also make Jones decide in accordance with Black's will. Black is going to make Jones decide to cheat only if he discovers that Jones is inclined not to cheat. Otherwise, Black will let Jones to make his own decision. It turns out that Jones decides to cheat on his own.

In the above scenario, it seems that when Jones is making the choice, he has no freedom to do otherwise than cheating. Despite this fact, however, we still have a strong intuition to think that Jones is morally responsible for what he actually chooses to do. A take-home message then is that the agent's access to alternative possibilities is not *necessary* for moral responsibility.[41] Since the free will concerned by most philosophers is the free will required by moral responsibility, the Frankfurt- Style cases have a further implication for free will—that the most important kind of free will does not consist of the ability to do otherwise. So here comes the tension between the Frankfurt-Style cases and compatibilism, particularly, the leeway approach to reasons-responsiveness. Two different ideas can be extracted from the Frankfurt-style cases. Accordingly, the Frankfurt-Style cases pose two challenges to the leeway approach to reasons-responsiveness.

[41] Frankfurt's original thought experiments and his seminal paper engender many following up debates. Say, whether the setting of the thought experiments can coherently rule out all the alternative possibilities; whether the thought experiments need to rule out all the alternative possibilities to establish the expected conclusion. For a collection of these debates, see McKenna and Widerker (2003)

i) The Challenge of Unnecessity

First of all, there is the direct intuitive judgment from the Frankfurt-style cases—the agent in question is morally responsible for his action even though he is robbed of the ability to do otherwise. This intuition implies that the ability to do otherwise is not required by moral responsibility and the relevant freedom. As shown above, reasons-responsiveness consists of the conditional ability to do otherwise (if there were sufficient reasons to do otherwise, she would recognize and respond to it). Thus, this intuition also implies that reasons-responsiveness is unnecessary for moral responsibility and freedom. Even though the Frankfurt-Style cases are usually taken as an argument against incompatibilism, it poses a challenge equally to incompatibilism and compatibilism. In the scenarios, the agents lack the ability to do otherwise in both the categorical sense and the conditional sense. Recently, Fischer (2012, 130) holds that the Frankfurt-Style cases pose an even more severe challenge to compatibilism than incompatibilism. For it is not so clear that the Frankfurt-Style cases have conclusively ruled out the categorical abilities to do otherwise because of the dilemma objection while it is clear that they have conclusively ruled out the conditional ability to do otherwise.[42] In a word, according to Fischer, there may be some disputes whether the Frankfurt-Style cases successfully show that the categorical ability to do otherwise is required by moral responsibility, there is no dispute that the Frankfurt-Style cases show that the conditional ability to do otherwise is not required by moral responsibility.

*ii)* The Challenge of Irrelevance

The second challenge goes deeper. Many compatibilists hold that the direct intuitive judgment from the Frankfurt-Style cases should be explained by a general philosophical principle, namely, the *actual-sequence view* of freedom, according to which, whether an agent is acting freely depends on the actual sequence which issues in the action. If the actual sequence is the right one, then the agent is free; and vice versa, the agent is not free. This actual sequence view implies the reasons-responsiveness is not only unnecessary but also irrelevant for moral responsibility.[43] Recall that the leeway approach involves a

[42] One of the important incompatibilist response is the dilemma objection: it is not clear how the counterfactual intervener can reliably predict the agent's choice or move unless determinism is presumed; if determinism is presumed, it will have the danger of begging the question against incompatibilism; if determinism is not presumed, then the agent still retain the ability to do otherwise. See Widerker (1995) and Kane (1996).

[43] The difference between a claim of unnecessity and of irrelevance can be put as follows. First, one can think that alternative possibilities are not necessary for moral responsibility, but still think that they are sufficient for moral responsibility. In this sense, even though alternative possibilities are not necessary, they are still relevant. Second, the irrelevance claim is about the idea that the alternative possibilities play no role in the consideration of control and freedom for moral responsibility; while the unnecessary claim does not commit to this strong idea. To illustrate, being familiar with Kant's works may not be a necessary condition for being a good philosopher. But it is a much

modal conception. According to this conception, reasons-responsiveness by nature is a modal property, which is characterized in terms of counterfactual scenarios/alternative possibilities. Typically, alternative possibilities are not part of the actual sequence. This contradicts the actual-sequence view, according to which we can tell whether the action is free or not free just by looking at the properties of the actual sequence.[44]

Saying that alternative possibilities are unnecessary for moral responsibility is a less committing claim than saying that alternative possibilities are irrelevant to moral responsibilities.[45] One can just accept the direct intuition from the Frankfurt-Style cases and reject the actual-sequence view. Accordingly, one can hold that the alternative possibilities are unnecessary but not irrelevant to moral responsibility. Unfortunately, the distinction is not well-recognized in the literature. In many situations, the two claims are conflated.[46] In the remainder of this chapter, I will just focus on the challenge of Unnecesity and will leave the second one to the next chapter.

### 1.4 A Split Within the Compatibilists

The Frankfurt-Style cases separate the compatibilists into two camps, as I described earlier – the first is the *leeway compatibilists* who insist on the conditional ability to do otherwise as essential to freedom and moral responsibility and provide more sophisticated conditional analysis of ability in order to handle the Frankfurt-Style cases; and the second is *source compatibilists*, who hold that freedom is the matter of the actual sequence. The contrast between the two kinds of compatibilists can be further put in the terminology introduced by Sartorio (2016), that they are providing different answers to the question of

stronger claim to say that being familiar with Kant's woks is irrelevant to being a good philosopher. Thus, a claim about irrelevance is stronger than a claim about unnecessity.

[44] Note that the source approach of reasons-responsiveness mentioned above is a kind of actual-sequence view, which tries to embed reasons-responsiveness into an actual-sequence framework. However, even though Fischer and Ravizza endorses an actual-sequence view, they still employ the modal conception of reasons-responsiveness (more on this point below). Thus, Fischer and Ravizza's account is still in tension with the Frankfurt-Style case. This tension has recently been flagged up by Sartorio (2015; 2016). She argues that this actual-sequence view implies a de-modalized conception of agency—free agency should be *exclusively* grounded in the actual sequence, rather than the alternative sequence. I will address this tension in the next chapter.

[45] See note 43 above.

[46] For example, it is a common expression made by compatibilists that the Frankfurt-Style cases show that alternative possibilities are not *required* for moral responsibility. (e.g., Fischer and Ravizza 1998, 30; McKenna 2015, 153) This claim is ambiguous because one can either interpret it in the weak sense that alternative possibilities are not necessary for moral responsibility or in the strong sense that they are not relevant to moral responsibility. Sartorio also faces this problem of ambiguity when she claims that the Frankfurt-Style cases show that moral responsibility is not *grounded in* alternative possibilities. One notable exception is Horgan (2015) who suggests a similar distinction of the inputs from the Frankfurt-Style cases.

what grounds agents' freedom.[47] The leeway compatibilists are providing the *Alternative Possibilities Answer* and the source compatibilists are providing the *Actual-Sequence Answer*.

> **The Alternative-Possibilities Answer:** When agents are free, their freedom is grounded, at least partly, in the fact that they are able to do otherwise.
>
> **The Actual-Sequences Answer:** When agents are free, their freedom is grounded only in facts pertaining to the actual processes or sequences of events issuing in their behaviour. (Sartorio 2016, 10)[48]

Though in Sartorio's view these two answers about freedom are competing, I will show that they are not if we distinguish freedom of agency and freedom of action. These two answers are to two different questions—what grounds freedom of agency and what grounds freedom of action. That is to say, the conflict between these two views can be explained away.

In this chapter, I aim to establish the view that reasons-responsiveness is leeway freedom which consists of the agent's ability to do otherwise. To reach this conclusion, I need to do two things. First, I will argue against the source approach to reasons-responsiveness. The target I have in mind is the mechanism-based reasons-responsiveness approach developed by Fischer and Ravizza (1998) and causal reasons-sensitivity approach developed by Sartorio (2015;2016). Second, I am going to find a way to reconcile the intuition from the Frankfurt-Style cases and the view that responsible agency consists of the conditional ability to do otherwise.

### 1.5 The New Dispositionalism

Before turning to the source compatibilists, I should devote some space to the new dispositionalists, which is the representative of the leeway compatibilists.[49] The phrase 'new dispositionalism' is coined by Clarke (2009) referring to a group of compatibilists – they draw the connection between dispositions and

[47] Though there are many controversies on the nature of grounding relation, this relation can be captured by the phrase 'in virtue of', such that A is grounded in B, then A obtains is in virtue of B. In addition, many philosophers think that grounding is a non-causal/metaphysical relation. It becomes more and more popular to view the free will problem as seeking the grounding conditions rather than merely sufficient and necessary conditions for freedom, e.g., Fischer 2018.

[48] Once again, a tricky point here is that some of the actual-sequence views of reasons-responsiveness retain the modal conception (e.g., Fischer and Ravizza 1998, McKenna 2013). These views involve an analysis with reference to the alternative possibilities, even though they are focusing on the property of the actual sequence rather than the ability of the agent.

[49] I think this move is necessary for one of my attack on Fischer and Ravizza's mechanism-based reasons-responsiveness have the danger of collapsing into an account which is similar to the new dispositionalism.

abilities and try to defend compatibilism with resources from the disposition literature. According to these views, agentive abilities share a similar metaphysical structure with dispositions. Thus, abilities can either be analyzed in a similar way as with dispositions (Smith 2003; Vihvelin 2004, 2013); or that they can be analyzed in terms of dispositions (Fara 2008). Specifically, the new dispositionalists learn an important lesson from the disposition literature that dispositions cannot be analysed in simple conditional because they can be masked or finked. Here is an example of simple conditional analysis for the fragility: the glass is fragile iff it will break when it is struck heavily. Here is an example of mask: the glass is covered by Styrofoam so that it will not break even if it is struck heavily. In this case, the glass retains fragility even though it does not satisfy the simple conditional. Here is an example of fink: A wizard nearby uses magic to remove the intrinsic property of the glass that accounts for its fragility (say, a specific molecule structure). However, whenever the glass is going to be struck, the wizard would immediately restore its property at the moment just before it is struck. In this case, the glass is not fragile even though it satisfies the simple conditional. The rationale is that dispositions are grounded in certain intrinsic causal properties, while the modal patterns characterized by the simple conditionals do not always track the intrinsic causal properties of dispositions: in the case of masks, the intrinsic causal properties are retained while the modal patterns do not obtain; in the cases of finks, the intrinsic causal dispositions are deprived of while the modal behaviour retains. Thus, the simple conditional analysis does not work. To analyze dispositions, more sophisticated conditionals are required to rule out the cases of finks and masks. Inspired by this lesson, the new dispositionalists want to revive the classical compatibilist ideas that freedom is the ability to do otherwise and that this ability is to be analyzed conditionally. They argue that the challenge of Unnecessity from the Frankfurt-Style cases can be avoided. This is because the setting of the neuroscientist does not remove the agent's ability to do otherwise; rather it is just a fink or a mask for that ability.[50] They propose a more sophisticated analysis of the ability to do otherwise where factors of the neuroscientist can be subtracted.

There have been many criticisms of the new dispositionalism—for example, there seems to be dissimilarities between dispositions and abilities;[51] and that the ability identified is too general to ground moral responsibility.[52] Here I want to flag up another drawback of this approach: in arguing for freedom as the ability to do otherwise, the new dispositionalists have to deny the intuition trigged by the Frankfurt-Style cases. Thus, they have to abandon a powerful resource against incompatibilism. In this chapter, I am

[50] Fara (2008) takes the setting of the counterfactual intervener as a mask. Vihevelin (2004) treats it as a fink. More recently, Vihevelin (2013) thinks that the setting of the counterfactual intervener can be either a mask (if the Frankfurt-Style case is presented in a "Bodyguard" way) or a fink (if the Frankfurt-Style case is presented in a "Preemptor" way).
[51] E.g., Vetter and Jaster 2017
[52] Clarke 2009; Whittle 2010

also defending the idea that freedom for moral responsibility, or more specifically, reasons-responsiveness, consists of the conditional ability to do otherwise while at the same time I will also endorse the intuition from the Frankfurt-Style cases. Namely, the intuition that the agent is morally responsible despite lacking the ability to do otherwise. Before presenting my proposal, I will first argue against two accounts which dissociate reasons-responsiveness with the ability to do otherwise.

## 2. Fischer and Ravizza's Mechanism-Based Reasons-Responsiveness Account

Fischer and Ravizza propose that the notion of reasons-responsiveness plays a significant role in grounding moral responsibility while they endorse the direct intuition from the Frankfurt-Style cases and hold that ability to do otherwise is not necessary for freedom and moral responsibility. That is, they face the challenge of Unnecessity. To solve this problem, they suggest switching from the *agent-based* reasons-responsiveness to the *mechanism-based* reasons-responsiveness. In their approach, reasons-responsiveness is no longer taken as an ability of the agent; rather, it is a modal property of the mechanism which issues in the action. Fischer and Ravizza provide little explanation for the notion of mechanism. Sometimes they just use other unclarified notions to explicate it, such as 'the process' or 'the way' that leads to the action.[53] As I will show below, the specification of this notion will become the Achilles' heels of their overall approach.

According to their approach, the agent is morally responsible for his action if the action is issued from the right actual sequence.[54] This further sets up two conditions: (i) the action is generated through a reasons-responsive mechanism and that (ii) the mechanism is the agent's own.[55] Now let's focus on the reasons-responsiveness condition. A mechanism's reasons-responsiveness can be characterized in the following way:

> Holding fixed the mechanism that actually issues in the action, if this mechanism operates in certain possible scenarios where there are sufficient reasons to do otherwise, the agent would recognize the sufficient reasons and react to the sufficient reasons.[56]

[53] Fischer and Ravizza 1998, 38.
[54] That is to say, Fischer and Ravizza also submit to the actual-sequence view of freedom motivated by the Frankfurt-Style cases.
[55] These two conditions are equivalent to the *guidance control*, a notion introduced by Fischer and Ravizza (1998) to capture the freedom which does not involve alternative possibilities, in contrast to the *regulative control* which involves alternative possibilities.
[56] This is a rough characterization without specifying and quantifying those possible scenarios. How to specify and quantify those possible scenarios depends on the degree of reasons-responsiveness required. Fischer and Ravizza propose that to adequately ground moral responsibility, the mechanism must be weak reasons-receptive (to

Similar to the leeway approach to reasons-responsiveness, Fischer and Ravizza's mechanism-based account analyzes the notion of reasons-responsiveness in terms of counterfactual scenarios. However, we have to bear in mind that what they aim to characterize is a property of the actual sequence, rather than an ability of the agent. Thus, their account abandons the agentive ability conception of reasons-responsiveness while retaining the modal conception. Suffice to say that this account still counts as a source view, though in what follows, I will show that it has the danger of slipping to a leeway view.

The focus on mechanism explains why in the Frankfurt-Style case the agent is morally responsible for the action despite lacking access to alternative responsibilities. Recall that in the Frankfurt-Style case, the neuroscientist, Black, does not intervene the agent's action, which means Black (and his device) is not part of the mechanism leading to the agent's decision and action. Thus, in evaluating whether the mechanism-based reasons-responsiveness is obtained, we should discount the neuroscientist and the device and only hold fixed the actual mechanism (say, the agent's deliberation which generates his action). Surely, this actual mechanism is reasons-responsive since, in the scenario, it has been stipulated that the agent is acting in a *normal* way.

The most important problem with the mechanism-based approach is that it seems to identify the wrong property to ground freedom and moral responsibility. This problem has been pointed out by Wallace, who thinks that the mechanism is the 'wrong locus' of moral responsibility. He writes:

> The deeper problem, however, lies in the supposition that questions of moral accountability can be clarified by attending exclusively to the modal properties of the "mechanisms" involved in action. This approach brings an objectifying, third-personal vocabulary to bear on phenomena that have their natural place within the deliberative perspective of practical reason, with the result that the intuitive locus of responsibility, the person, seems to drop out of view. (Wallace 1997, 159)

Under one interpretation Wallace seems to raise the problem of disappearing agency which plagues the naturalist account of agency—that is, a naturalistic account of agency (e.g., the causal theory of action) cannot capture some essential aspects of our conception of agency. At least Fischer interprets the criticism in this way—he suggests that his mechanism-based account is compatible with, but is not necessarily equivalent to, the naturalistic account. Meanwhile, he holds there are no particular reasons to oppose a naturalistic account of agency. Besides, he thinks that the agent does not drop out of the view because "it is the person who is morally responsible (partly) in virtue of the operation of his own suitably reasons-

recognize the reasons in at least *one* possible scenarios) and strong reasons-reactive (to react the reasons in *some* possible scenarios).

responsive mechanism".[57] I agree with Fischer that the agent would not simply "drop out of the view" by looking at the mechanism which brings about the agent's action. I also think that the problem of disappearing agency is not quite relevant to the mechanism-based account and can be set aside.[58] However, the issue raised by Wallace is more subtle than Fischer has thought: when we are speaking of responsible agency, we are speaking of certain agentive properties to ground moral responsibility. The mechanism-based reasons-responsiveness, however, is not an agentive property directly attributed to the agent. Thus, it is not clear that it retains the grounding power for freedom and moral responsibility.

To see how this conclusion is reached, we should examine carefully the relation between the agent and his mechanism. Fischer and Ravizza say very little on the notion of mechanism, let alone its relation to the agent. What we only know is that the agent is morally responsible (partly) in virtue of his mechanism's being reasons-responsive. There are two possible interpretations: it may be that the mechanism operates at the sub-personal level. In this case, it is the agent who responds to the reason while the agent's reasons-responsiveness is subserved by the low-level property, namely, mechanism-based reasons-responsiveness. Or it may be that the mechanism operates at the personal level. In this case, the mechanism, as a representative part of the agent, responds to reasons.[59] In what follows, I will argue that in the first case the problem identified by Wallace recurs; while in the second case, that problem can be avoided but new problems are triggered.

### 2.1 The mechanism operates at a sub-personal level

If the mechanism operates at a sub-personal level, then mechanism cannot literally respond to reasons because talking of reasons only makes sense at a personal level; rather, it is the agent who responds to reasons while the mechanism is just the underlying condition which enables the agent to respond to reasons. Understood as a sub-personal property, the mechanism-based reasons-responsiveness is not the reasons-responsiveness as we ordinarily understand it.

---

[57] Fischer 2012, 158.

[58] I will discuss this problem in detail in Chapter 5 and Chapter 6.

[59] For either reading, we can find some textual clue. For the first reading: in Fischer and Ravizza's characterization of mechanism-based reasons-responsiveness, it is the agent rather than the mechanism who recognizes and reacts to reasons. The role of the mechanism is just a set of underlying conditions which enable the agent to respond to reasons (Fischer and Ravizza 1998, 41); in addition, sometimes they suggest that the mechanism-based reasons-responsiveness is a dispositional/modal property attributed to the actual sequence (Fischer and Ravizza 1998, 53). For the second reading: they sometimes refer to the actual normal mechanism as the 'deliberative mechanism'. (Fischer and Ravizza 1998, 36) It seems natural to say that a deliberative mechanism is operates at the personal level and literally responds to reason.

Here is an analogy. In philosophy of mind, it is widely presumed that high-level mental properties are underpinned by low-level physiological or physical properties. For example, the feeling of pain is underpinned by C-fiber firing. This connection is probably contingent rather than conceptual. For there is the so-called *explanatory gap* between the mental properties and the physical properties.[60] We can imagine that being pain is underpinned by other physiological properties when it comes to other creatures or in other possible worlds where the laws of nature are different. Simply put, being pain is just nomologically dependent on C-fiber firing. Unless the explanatory gap between the mind and the body can be filled, the conceptual connection between mental and physical cannot be established. Accordingly, a property which is conceptually connected to being pain will probably not also be conceptually connected to being C-fiber firing. For example, there is a conceptual connection between being pain and being uncomfortable. But probably, there is no conceptual connection between being uncomfortable and C-fiber firing.

This mental-physical relation characterized above is analogous to the agent-mechanism relation. Just like pain is underpinned by C-fiber firing, the agent's responding to reason, as a high-level property, is underpinned by the mechanism-based reasons-responsiveness which is a low-level property. Accordingly, the connection between the agent-based reasons-responsiveness and the mechanism-based reasons-responsiveness is better regarded as contingent rather than conceptual. Understood as a sub-personal property, the mechanism-based reasons-responsiveness is no longer the reasons-responsiveness in its original sense. Rather, it is just a place holder for any sub-personal properties which enable the agent to respond to reasons in the actual world. Unless in a metaphorical sense, a sub-personal mechanism never reacts to or recognizes reasons. (Just like the C-fibers can fire but can never feel pain. Otherwise, it would commit the Category Mistake.) Arguably, there is a similar explanatory gap between the agents' responding to reasons and the low-level mechanism-based reasons-responsiveness. The agent's responding to reasons is just nomologically dependent on the mechanism-based reasons-responsiveness which is a sub-personal property.

If there is a conceptual gap between agent-based reasons-responsiveness and mechanism-based reasons-responsiveness, then the tight conceptual connection between agent-based reasons-responsiveness and responsibility does not transit to mechanism-based reasons-responsiveness and responsibility. To illustrate, recall that being pain is conceptually connected to being uncomfortable, but C-fiber firing is not. Analogously, there is a close conceptual connection between (agent-based) reasons-responsiveness and moral responsibility, but such connection probably will not obtain when switching to the mechanism-

[60] The term 'explanatory gap' is due to Levine (1983).

based approach (where the mechanism is construed as a sub-personal mechanism). The loss of this conceptual connection is fatal to the mechanism-based approach. In particular, philosophers including Fischer recently emphasize that free will debate surrounding moral responsibility is not just about seeking the necessary control condition for moral responsibility, but also answering the question of what grounds moral responsibility.[61] Usually, a grounding relation is taken as a non-causal explanatory relation. Thus, if A is grounded in B, then there must be a certain conceptual connection between A and B such that A is non-causally explained by B. It is an attractive idea to take moral responsibility is grounded in reasons-responsiveness because of the close conceptual connection.[62] Switching to the sub-personal mechanism, however, this conceptual link no longer maintains. Now I am at the position to reformulate Wallace's worry: the mechanism-based reasons-responsiveness, understood as a sub-personal property, does not hold the conceptual connection to moral responsibility. Consequently, it is difficult to see how it can ground moral responsibility.[63] Of course, there is still a way to circumvent this criticism. That is, to understand the mechanism as one operating at the personal level but as a subpart of the agent. This is the approach I now turn to.

### 2.2 The mechanism operates at the personal level

Unlike a sub-personal mechanism which underlies the agent's responding to reasons, a personal-level mechanism can literally respond to reasons as a representative part of the agent. To illustrate this relation, here is another analogy. The University, say, is conceptually composed of certain sub-institutions, such as departments, faculties and research groups. Let's suppose that at the University, there is a research group A, which has the leading scientists and advanced experimental devices. The research group A is able to crack a very significant scientific problem. In this case, we can claim that the University is able to crack a scientific problem *in virtue of* the fact that the research group A is able to crack that scientific problem. Here, the research group A is a conceptual part of the University and can be representative of the University with respect to cracking the scientific problem. This analogy can mirror the Frankfurt-Style case by setting up a situation in even though the research group A is able to crack the scientific problem, the University as a whole is not able to do so. Just suppose that in the University, there is a research group

[61] For the emphasis on the grounding role of free agency, see Sartorio (2016, chapt. one) and Fischer (2018).
[62] Note that almost everyone (incompatibilists and compatibilists included) agree that reasons-responsiveness is an important and necessary condition for moral responsibility, what they disagree is whether it is sufficient.
[63] I am not claiming that mechanism-based reasons-responsiveness, understood as the properties of sub-personal mechanism, cannot ground moral responsibility. Rather, my point is that we cannot attribute the grounding power to it as easily as to the agent-based reasons-responsiveness. In this sense, switching to the mechanism makes the reasons-responsiveness theory lose its original intuitive appeal.

B which is nasty and despicable—whenever the research group A is going to make the relevant result, the research group B will stealthily sabotage the data and prevent the group A from getting the results. In such a situation, the research group A, *on its own*, is able to crack the scientific problem,[64] while the University, considered as a whole, is not for the effort from one part, namely, the research group A, is always counteracted by another part.

Now apply this analogy to the agent/mechanism dichotomy. Just like the University is *conceptually* composed of certain sub-institutions, the agent is *conceptually* composed of certain mechanisms. According to a recent clarification of Fischer, in the Frankfurt-Style cases, the agent is composed of "actual-sequence mechanism" plus the "alternative-sequence mechanism(s)", namely, the potential manipulation of the neuroscientist.[65] When the nasty research group B is subtracted, research group A can crack the scientific problem on behalf of the University. Similarly, when the nefarious neuroscientist is subtracted, the normal mechanism can respond to reasons on behalf of the agent.

In this reading, there is no difference in meaning between the agent-based reasons-responsiveness and mechanism-based reasons-responsiveness. Both notions refer to the ability to respond to reasons.[66] The only difference is that the former is attributed to the whole agent, while the latter only to a part of the agent, namely, the mechanism in operation. Thus, a close conceptual connection between reasons-responsiveness and moral responsibility still obtains when switching to the mechanism. Just as we can say that the University deserves praise in virtue of its research group A's scientific achievement, we can also claim that the agent is morally responsible in virtue of his mechanism is reasons-responsive. Thus, this reading helps to avoid Wallace' worry because a personal-level mechanism of the agent can serve as a proper locus of moral responsibility.

Now let's consider some problems in this reading.

---

[64] Given the presence of the nefarious research group B, some may be reluctant to accept the claim that the research group A is still able to crack the scientific problem. But at least in a sense it is still able to do it. For example, William is a normal healthy person but unfortunately his limbs are currently bound with rope. Is he able to walk? In a sense he can but in another he cannot. In the philosophical discussion of ability, is common two distinguish two different kinds of ability. For example, Vihvelin (2013) distinguishes the narrow ability and wide ability. Narrow ability refers to the agent's intrinsic capacity to perform the action; wide ability refers to the agent's intrinsic capacity plus the extrinsic factors which provides the chance to perform the action. Thus, if an agent is confronting a task just as William, then he loses the wide ability while remain the narrow one. Similarly, in the above case, we can say that the research group A has the narrow ability to crack the scientific problem even though it does not have the wide ability to do so.

[65] Fischer 2012, 157.

[66] Compare the sub-personal mechanism which cannot literally respond to reasons.

### 2.3 The Problem of Individualization of Mechanism and *Ad Hoc Construction*

The first problem is about the individuation of mechanisms. According to Fischer and Ravizza, to assess whether the operating mechanism is a reasons-responsive one, we need to hold the mechanism fixed, and make certain modal inferences to see whether it will recognize and react to reasons in counterfactual scenarios. Thus, the individuation of mechanism is important to the approach. Is it possible to provide a principle to identify the pertinent mechanism?

Recently, McKenna (2013) convincingly argues that there is no principled way to individualize a mechanism because any operating system is further composed of sub-systems. Even worse, the distinction of agent/mechanism is not stable. Just imagine a camera that is sensitive to multiple light conditions. This sensitivity is explained by different systems that operate at different light conditions. For example, the sub-system A only operates on condition-A and sub-system B operates on condition-B. Now the question is, in assessing the sensitivity of the Camera, which mechanism shall we hold fixed? If we hold fixed a specific sub-system, then the mechanism is surely not light-sensitive for it only operates on a specific light condition. However, when we hold fixed more sub-systems as the mechanism, then the camera/mechanism distinction tends to disappear. Now the agent/mechanism distinction faces a similar dilemma. On the first horn, if we take what is actually in operation as the pertinent mechanism, then what we identify may just be a sub-system of the whole agentive system. This sub-system is too narrow to be responsive to a wide range of reasons, and then it is inadequate to ground moral responsibility; On the second horn, if we want to make the mechanism more inclusive, then the mechanism identified will be 'functionally equivalent to the agent', as McKenna argues. There is no principled way to prevent this slippery slope so that there is no way to make the distinction between the agent and mechanism.[67]

Is this problem fatal to the mechanism-based approach? At least Fischer and Ravizza do not think so. They admit that a lack of a principled way to individualize the mechanism may somehow be a disadvantage to their approach, but they insist that we still can appeal to intuitions to distinguish a mechanism from another. Now the dialectical situation is like the this: we do not have any independent principle to individualize the mechanism for actions so we do not know how to accurately identify the pertinent mechanism for responsibility ascription. However, at least in the Frankfurt-Style cases, we can easily tell the difference between the mechanism operates in the actual scenario and the one controlled by the neuroscientist in counterfactual scenarios. Thus, Fischer and Ravizza seem to have resources to claim

[67] Note that McKenna's argument only runs when the mechanism is understood as operating at personal level. This is because, if a mechanism is a sub-personal one, then no matter how wide we take it to be, the agent/mechanism distinction still holds.

that in the Frankfurt-Style case, the agent is not reasons-responsive while the operating mechanism is reasons-responsive and this grounds the agent's moral responsibility.

The concern is that lacking a way to accurately identify the pertinent mechanism, the notion of 'mechanism' is just an *ad hoc* construct to salvage the reasons-responsiveness theory from the Frankfurt-Style cases. For all the theoretical work that the notion of mechanism does is just to provide a license for discounting the setting of the neuroscientist in the counterfactual inferences when assessing the pertinent reasons-responsiveness which grounds moral responsibility. Thus, after denying the old subject, namely the agent, to be reasons-responsive in the Frankfurt-style cases, Fischer and Ravizza invent a new subject, namely the 'mechanism', to bear the property of reasons-responsiveness which grounds moral responsibility, even though they provide little articulation about what the mechanism actually is. And this point related to another problem which I now turn to.

### 2.4 An ability to do otherwise in disguise?

Recall that under this reading, there is no difference in meaning between the agent-based reasons-responsiveness and the mechanism-based reasons-responsiveness. Since the agent-based reasons-responsiveness consists of the ability to do otherwise, it then follows that the mechanism-based reasons-responsiveness is just an ability to do otherwise in disguise. In other words, Fischer and Ravizza's mechanism-based approach collapsed into a leeway compatibilist approach. Another way to see this point is to compare the mechanism-based approach and the new dispositionalism approach. According to the new dispositionalism, responsible agency consists of the ability to do otherwise. If it turns out that there is no substantial difference between these two approaches, then it indicates that the mechanism-based reasons-responsiveness is just an ability to do otherwise.

The new dispositionalism and the mechanism-based approach have seeming different explanations for why the agent is responsible for his action in the Frankfurt-Style cases. According to the new dispositionalism, the relevant freedom, namely, the agent's ability to do otherwise is retained even in the Frankfurt-Style case because the intrinsic properties which subserve the agent's ability are intact. In addition, this freedom is analyzed in term of possible scenarios where the subject of that freedom (the agent) decides and act differently. In this modal analysis, the factor of the neuroscientists should be screened off for it is taken as a fink or a mask.

Alternatively, according to the mechanism-based approach, the relevant freedom, namely, the mechanism-based reasons-responsiveness is retained even in the Frankfurt-Style case because the actual

mechanism is not affected by the neuroscientist. Besides, the relevant freedom is analyzed in terms of possible scenarios where the subject of freedom (the mechanism) recognizes and reacts to reasons differently.[68] In this modal analysis, the factor of the neuroscientist should be screened off because it is not part of the actual mechanism.

Admittedly, there are differences between these two approaches: :1) they assign freedom to different subjects—for the new dispositionalists, it is the agent who bears the freedom; for Fischer and Ravizza, it is the agent-like mechanism. 2) they provide different stories about why the factor of neuroscientist should be disregarded in the modal analysis of freedom. However, the similarities between these two approaches are significant. Both take that there is a kind of freedom retained in the Frankfurt-Style cases. Moreover, they analyze the relevant freedom in a similar modal way—the neuroscientist should be screened off in the counterfactual inferences. [69]

In summary, Fischer and Ravizza try to dissociate reasons-responsiveness with the ability to do otherwise by introducing the notion of mechanism-based reasons-responsiveness. However, their approach faces a dilemma: if the mechanism operates on the sub-personal level, then the mechanism-responsiveness lose the original grounding power for freedom and moral responsibility; if the mechanism operates on the personal level, then the notion of mechanism becomes an *ad hoc* construction and the mechanism-based reasons-responsiveness becomes an ability to do otherwise in disguise.

### 3. Sartorio's Causal Reasons-Sensitivity Account

Recall that Sartorio distinguishes two competing theses regarding the question of what grounds free agency:

> **The Alternative-Possibilities Answer:** When agents are free, their freedom is grounded, at least partly, in the fact that they are able to do otherwise.

[68] Note that this only applies to the personal-level mechanism. If a mechanism is operating in sub-personal level, it is the agent not the mechanism that responds to the reasons. Thus, the similarity only obtains when the mechanism is understood as in personal level.

[69] Recently, Franklin (2015) also argues that Fischer and Vihvelin's dispute regarding the ability to do otherwise is just a verbal one. My point differs with Franklin in several respects: First, we can only conclude that there is no substantial dispute between Fischer and Vihvelin (and other new dispositionalists) when the notion of mechanism is understood as in the personal level. In addition, I disagree with Franklin's reconstruction of Fischer's view. Specifically, Franklin suggests that Fischer will agree that the agent's conditional ability to do otherwise is not ruled out by the Frankfurt-style cases. However, Fischer (2012, 130) put it explicitly that the Frankfurt-style cases divisively rules out the agent's conditional ability to do otherwise.

> **The Actual-Sequences Answer:** When agents are free, their freedom is grounded only in facts pertaining to the actual processes or sequences of events issuing in their behaviour.

The leeway compatibilists take the first answer for they think that freedom is the ability to do otherwise, which is cashed out in terms of alternative possibilities; the source compatibilists take the second answer, for they think that the Frankfurt-Style cases have convincingly shown that the agent is acting freely if the action is generated through a right actual sequence. Like Fischer and Ravizza, Sartorio is siding with the source compatibilists. In addition, Sartorio agrees with Fischer and Ravizza that reasons-responsiveness is the control condition for moral responsibility. However, Sartorio is not satisfied with Fischer and Ravizza's mechanism-based account for it retains the modal conception of reasons-responsiveness; that is, reasons-responsiveness is still analyzed in terms of possible scenarios. She argues that the modal conception contradicts the Actual-Sequence Answer, according to which freedom is grounded exclusively on the actual sequence.[70] The main motivation for Sartorio's account is to resolve the seeming inconsistency which plagues the mainstream reasons-responsiveness theory—on the one hand, the reasons-responsiveness is a modal concept to be articulated in terms of possible scenarios; on the other, the Frankfurt-Style cases motivates an actual-sequence conception of free agency, according to which freedom grounds on the actual sequence. To solve this tension, Sartorio develops Causal Reasons-Sensitivity account, according to which the notions of reasons-sensitivity are not understood "in counterfactual or modal terms, but exclusively in terms of actual causal histories".[71]

The development of the account consists of two steps. In the first step, Sartorio proposes that absence enters into the causal relation. A typical example of absence causation is that one's forgetting to water the flower causes the flower to wither. The next move is to contend that in acting for reasons, the agent's action is an outcome of both reasons and absences of reason. For example, If I choose to stay at home for the weekend, one of the explanations for the choice is that I want to play some video games. But the lack of interesting movies on the scene can also figure in the explanation. That is to say, my staying at home is an act of responding to both reasons and absences of reasons. Combine these two theses, we have a new picture of agency—both reasons and absences of reasons can be elements of the actual sequence which brings out the agent's action. Sartorio refers to this conception of agency as *Causal Reasons-Sensitivity* (CSR), which is characterized in the following way:

[70] More on this point in Chapter 3.
[71] Sartorio 2016, 123

> CRS: An agent is reasons-sensitive in acting in a certain way when the agent acts on the basis of, perhaps in addition to the *presence* of reasons to act in the relevant way, the *absence* of sufficient reasons must cause the act.[72]

Note that unlike the notion of reasons-responsiveness which is to be analyzed in modal terms, the notion of reasons-sensitivity is characterized only in terms of facts about causal history. In other words, the CSR is a de-modalized account of rational ability.

Sartorio further argues that causal reasons-sensitivity is the basic freedom which can ground moral responsibility. To illustrate, imagine two persons who are using drugs—one is a non-addict and the other is an addict. The causal reasons-sensitivity account can explain the difference of freedom and moral responsibility between them: when using drugs, the non-addict is acting freely and responsibly in the sense that he is responding both to reasons (e.g., the fact that he has a desire to get high) and absence of reasons (e.g., the absence of the fact that one is paying him 500 pounds to quit using drugs); while the addict is not acting freely or responsibly. This is because even if the addict may be responding to the same reason as the non-addict, he is not responding to those absences of reasons.

With this account, Sartorio can explain how the agent in the Frankfurt-Style case retains his freedom for moral responsibility. Recall that the agent (Jones) is not able to do otherwise than cheating on the exam because of the neuroscientist's setting. Neither he is reasons-responsive for reasons-responsiveness consists of the ability to do otherwise. However, Jones is still acting in *reasons-sensitive* way when he chooses on his own. This is because, when he is choosing, he is responding both to reasons and particularly, absences of reasons. To repeat, being reasons-sensitivity is just a matter of being caused in a specific way. We can imagine that Jones chooses to cheat for the reason that he really wants to pass as well as for the absence of serious potential punishment that he would face if getting caught. From the perspective of the actual-sequence view, the causal history leading to the decision is constituted with a combination of reasons and absence of reasons—it is a *right* actual sequence which grounds the agent's moral responsibility. Therefore, Sartorio provides a solution for the apparent tensions between the reasons-responsiveness and the Frankfurt-Style cases. Recall that the Frankfurt-Style cases post two different challenges to the traditional reasons-responsiveness account—first, in the Frankfurt-Style case, the agent is not able to do otherwise and *ipso facto* not reasons-responsive but still morally responsible for his action, which implies that reasons-responsiveness is unnecessary for moral responsibility; and second, the Frankfurt-Style cases motivate an actual-sequence view, according to which responsible agency can be articulated with features of the actual sequence and without the features of alternative possibilities,

[72] Sartorio 2016, p. 133

which implies that reasons-responsiveness is irrelevant to moral responsibility. Sartorio's causal reasons-sensitivity account handles these two challenges together. First, according to her account, being reasons-sensitive does not require even the conditional freedom to do otherwise; and second, being reasons-sensitive is all about the matter of the actual causal history.

### 3.1 The Problems with the Causal Reasons-Sensitivity Account

If absences can enter in the actual causal chain of the action, then there seem to be infinite numbers of absences in the actual causal chain. The absences are *explosive.* For example, I go to work as usual for the reasons that I do not want to be fired, also for the absence of reasons, say being sick. However, I can also say that I go to work for the absence of an attack by the aliens. Sartorio realizes this problem but does not say much about how to handle it. In some passages, she seems to think that we can rely on our intuitions to pick up the absences which are "highly relevant to the agent's exercise of his capacity to act for reasons" and rule out "less potentially significant absences".[73] I have doubts about whether our intuitions are adequate to identify the actual sequence which is composed of reasons and absences of reasons. However, my main concern is not about the metaphysics of absence causation. Rather, it is about the approach to take free agency being exclusively grounded in the actual causal history—this approach, by nature, is past-directed and de-modalized. It fails to capture two important features of free agency (or the freedoms for the agent).[74] The first one is *future-orientation* and the second one is *modality*.

First, our beliefs about our agency are mostly *future-directed*.[75] Agency is mostly about choosing and acting, while choosing and acting are about making a difference to the future, not to the past. Accordingly, when we are ascribing free agency to a particular agent, most of our concerns is how he or she will behave in the future and the relevant impacts that it may have on the future. That is to say, we need to adopt a forward-looking stance in assessing free agency. If, as Sartorio argues, free agency is exclusively grounded in the causal history, then it is entirely a matter of the past. It strikes me as a very strange idea that whether an agent is *currently* free exclusively depends on the past, namely, how his actions *came* about.

Now let's focus on another feature of Sartorio's account—that the analysis of free agency should avoid making references to modal terms such as alternative possibilities. However, free agency is essentially a modal concept. The talk of it essentially involves modal considerations. Gilbert Ryle's discussion on

[73] Sartorio 2015, 110.
[74] As I will show, the choice of terminology at this point is important.
[75] For this point, see also Vihvelin 2018.

agentive predicates is helpful here. ln *The Concept of Mind,* Ryle introduces the notion of *semi-dispositional and semi-episodic* to describe a range of predicates which are related to human agency, such as 'ready', 'on guard', 'careful', 'on the look out', and 'resolute'. These words are semi-dispositional and semi-episodic because they "do not signify the concomitant occurrence of extra but internal operations, nor mere capacities and tendencies to perform further operations, but something between the two."[76] He uses the word 'careful' as an example to illustrate the idea:

> The careful driver is not actually imagining or planning for all of the countless contingencies that might crop up; nor is he merely competent to recognize and cope with any one of them, if it should arise. He has not foreseen the runaway donkey, yet he is not unprepared for it. His readiness to cope with such emergencies would show itself in the operations he would perform, if they were to occur. But it also actually does show itself by the ways in which he converses and handles his controls even when nothing critical is taking place. (Ryle 1949/1990, 47)

The idea from the passage is that being careful is not just a matter of what the agent has done or is doing (e.g., paying attention to the road), it is also about what the agent was about to do in a series of hypothetical situations (e.g., bypassing a donkey on the way). I suggest that being free has a similar feature. When judging whether an agent is free, we are of course considering specific episodic properties which the agent is instantiating currently, (say, being conscious, being in normal physiological and psychological conditions, being away from physical constraints). However, we are also considering the behaviours and actions of the agent in a range of hypothetical scenarios. Thus, to evaluate whether the agent is free or whether the agent possesses free agency, we need to make a series of modal inferences, asking questions such as how the agent will act if certain situations are presented. It seems that a complete account of agency should do justice to both the episodic dimension and the modal/dispositional dimension.

Another reason for the claim that free agency essentially involves modal considerations is that freedom of agency is constituted with certain agentive abilities. Here the notion of abilities is not restricted to 'the abilities to do otherwise'; rather it can be abilities in a more mundane sense (say, I am able to speak Cantonese or she is able to swim) which does not directly trigger the metaphysical concerns in the free will debate. There are two main proposals for the semantics of abilities.[77] The first one is to reduce the claims of abilities into claims of conditionals, which I have touched upon in section 1. The second one is

[76] Ryle 1949/1990, 47
[77] For a review of the ability literature, see Maier 2018.

the Lewis/Kratzer proposal which reduces the talk of abilities to the talk of possibilities. That is, being able to X can be analysed in terms of possibilities of X-ing under certain conditions. Here is the often-quoted passage from David Lewis:

> To say that something can happen means that its happening is compossible with certain facts. Which facts? That is determined, but sometimes not determined well enough, by context…. It is likewise possible to equivocate about whether it is possible for me to [X], or whether I am able to, or whether I have the ability or capacity or power or potentiality to. Our many words for much the same thing are little help since they do not seem to correspond to different fixed delineations of the relevant facts. (Lewis 1976, 150)

For example, when we say that Chuang is able to play the piano, in one particular sense, we just mean that it is possible for Chuang to play the piano given that he has had the relevant skill and competence; in another sense, we mean that it is possible for Chuang to play the piano given that he has the relevant skill and competence plus that there is a piano accessible to him. According to this proposal, there is a conceptual correlation between the agent's ability to X and the possibility of the agent's X-ing under certain conditions, while what conditions should be taken into account is a matter of contexts.[78] Both the conditional proposal and the possibilities proposal point to the connection between abilities ascription and modal inferences: to judge whether an agent possesses a certain ability, we need to consider how the agent act in a series of possible situations. Thus, if free agency is constituted with certain abilities, say, the ability to recognize reasons and to act in light of reasons, it is implausible to view it as a property *exclusively* grounded in the actual sequence. Embedding the abilities to respond to reasons into this actual causal framework implies that an agent is able to respond to reasons only when he is *actually* responding to reasons. This is the Megarian view of ability, according to which an agent is able to X *only* when he is X-ing. This contradicts the common conception of abilities such that abilities can be ascribed to the agents even when the agents are not exercising those abilities.[79]

### 3.2 Two Models of Freedom: Competing or Complementary?

Some may find my criticisms above unfair, for the actual-sequence view and the alternative-possibilities view have different focuses in analyzing freedom: the former focuses on the past and the causal structure, and the latter focuses on the future and the modal structure. That is to say, the actual-sequence view

[78] See also Kratzer 1977.
[79] The rejection of this view traces back to Aristotle. For a discussion, see Maier 2018.

allows us to assess agency from a de-modalized and past-directed perspective. Thus, to rebut Sartorio's account, I need to show that the alternative-possibilities view is superior to the actual-sequence view in analyzing freedom.

Here is my clarification: I am not rebutting the actual-sequence view *simpliciter*; my claim is only that the actual-sequence view is not the proper framework for *freedom of agency* (and in this case, reasons-responsiveness); while it may be the proper framework for *freedom of action*. The difference between the alternative-possibilities view and the actual-sequence view leads to an assumption that these two models are providing competing answers for the same question—what grounds freedom.[80] However, as I will show, the two models actually aim to answer different questions—the alternative-possibilities view aims to answer the question what grounds free agency; while the actual-sequence view aims to answer the question what grounds free action. We tend to take the actual-sequence view and the alternative-possibilities view be two competing views of free agency, mainly because of the ambiguity of the "free" predicate.

In the free will literature, the "free" predicate is used in two different ways. First, it can be attributed to the agent, e.g., "the agent is free", or "the agent possesses free agency". Second, 'free' can also be attributed to actions (or sometimes choices and decisions): we can say that 'the action is free' or that 'the agent acts freely'. To avoid unnecessary ambiguities, I suggest that it is useful to reserve 'free agency' to the first usage; and 'free action' to the second usage. The question is, whether such a linguistic difference indicates a metaphysical difference. Some may hold a negative answer because they think that being a free agent is a sufficient and necessary condition for the agent to perform a free action. This means that the same metaphysical criteria are employed to assess freedom of action and freedom of agency. Contra this view, I propose that there is a significant metaphysical difference between free action and free agency. And because of that, when we are assessing free actions and free agents, different factors are taken into considerations.

First of all, the notions of agent and action, though bear a closed conceptual relationship, belong to different ontological categories. Agents are substance-like particulars, which tend to remain generally stable within a certain period of time. By comparison, actions are temporal particulars, which are datable, locating at spatial-temporal points and usually associated with changes. Given that agents and actions are of different categories, it is natural to presume that the 'free' predicates, being attributed to these two subjects, denote different properties.

[80] E.g., Sartorio 2016.

Besides, I propose that free agency should be better taken as a set of abilities possessed by the agent. That is, to ascribe freedom to the agent is to ascribe a set of abilities to the agent.[81] This ability view explains why freedom of agency is future-directed and has a modality dimension: first, the ascription of abilities to the agent is partly based on the record of past performance, and partly based on our expectation about how the agent will perform in future situations. Moreover, abilities involve modal thinking. When we are ascribing abilities to the agent, we need to make a series of modal inferences.

Now let's turn to the notion of 'free action'. In one sense, a free action is an action performed by a free agent. Accordingly, the meaning of 'free action' is derived from the meaning of 'free agent'. However, there seems to be a more idiosyncratic sense of 'free action'. In particular, when we say that the 'action is free', we are not focusing on the modal facts of the agent such as how she would act in possible scenarios. Rather, we are speaking of the specific process through which the action comes about: if the action is generated through a right process, then the action is free; or vice versa, if the action comes about through a wrong process, then the action is unfree. Note that this conception of free action is neutral to most parts of the free will debate. Different philosophers set up different criteria for the *right* process through which a free action is generated. For example, a compatibilist may propose that for an action to be free, the action should result from a deliberational mechanism (e.g., Fischer and Ravizza 1998) or from a harmonious psychological structure (e.g., Frankfurt 1971; Watson 1975); an incompatibilist may propose that a free action is produced through an indeterministic event sequence (e.g., Kane 1996); or an agent-causalist may contend that free action is caused by the agent as an irreducible substance (e.g., Pereboom 2001). Set aside the variations, all of these proposals specify free action with reference to the processes or the causal histories through which the actions come about.

To summarize, the difference between freedom of action and freedom of agency is not just a linguistic one. It indicates a difference of considerations in the assessments. In assessing free agency, we need to make modal inferences and usually adopt a forward-looking perspective.[82] By comparison, to evaluate free actions, we usually adopt a backwards-looking perspective, focusing the properties of the actual

[81] The idea that free agency is about possessing certain abilities to act can be supported by the fact that the abilities (to do otherwise) are central to the debate between compatibilists and incompatibilists. Both compatibilists and incompatibilists agree that to be a free agent is to have certain abilities to act in an appropriate sense. The disagreement is on how to analyse these abilities.

[82] Some may suggest that free agency may also involve a historical and backward-looking dimension. Specifically, many philosophers working in free will and moral responsibility have proposed the so-called historical conditions for moral responsibility (e.g., Kane 1996; Fischer and Ravizza 1998; Pereboom 2001). These historical conditions concern about how the characters and personalities are formed. I am neutral to the question whether moral responsibility require these historical conditions. However, in this chapter, I will reserve 'free agency' to refer to the agentive abilities.

causal history which lead to the action. Accordingly, we are also allowed to suspend certain modal considerations such as how the agent will act or would act in counterfactual situations. Thus, unlike in the assessment of free agency, we can adopt de-modalized perspective in the assessment of free action.[83]

| | **Attributed to** | **Focus on** | **Considerations in Assessment** |
|---|---|---|---|
| **Leeway Freedom** | Agent | The abilities of the agent | Modal structure; Future-Orientation |
| **Source Freedom** | Action | The way which a specific action comes about | Causal Structure; Past-Orientation |

*Table 1*

If the distinction made above is substantial, then the alternative-possibilities view and the actual-sequence view aim to answer different questions about freedom: the alternative-possibilities view is about free agency, while the actual-sequence view about free action. These two views need not be competing, rather, they can be complementary. Of course, the tension between the alternative-possibilities view and the actual-sequence view (and correspondingly, the tension between leeway compatibilism and source compatibilism) is mainly due to the Frankfurt-Style cases. To relieve the tension, I need to explain why the agent in the Frankfurt-Style cases is morally responsible despite not being able to do otherwise. In the next section, I will provide an explanation, which hinges on the distinction between freedom of agency and freedom of action.

## 4. How to Deal with the Frankfurt-Style Cases?

We have seen two compatibilist reactions to the Frankfurt-Style cases. Source compatibilists endorse the direct intuitive judgment elicited by these cases and contend that the ability to do otherwise is non-necessary and irrelevant to the freedom required by moral responsibility. This move is taken by Fischer & Ravizza and Sartorio and we have seen the problems they confront. By comparison, leeway compatibilists (particularly, the new dispositionalists) hold that the Frankfurt-style cases fail to show that the ability to do otherwise is not necessary for freedom. The problem with it is that leeway compatibilists have to deny a very strong intuition from the Frankfurt-Style cases and give up a powerful weapon against

[83] This is not to say that the assessment of free action does not involve any modal thinking. Even if we grant that whether an action is free depends on its causal history, we may still require modal analysis for the causal relation in question. And as I will show in Chapter 4, the notion of control involves both actual and modal dimensions.

incompatibilism. However, the compatibilist geography is not exhausted by these two reactions. In what follows, I want to show a way out—that is, we can, on the one hand, accept the direct intuitive judgment of the Frankfurt-Style cases; on the other hand, we can accept that ability to do otherwise is necessary for responsible agency. This proposal is hinged on the distinction between free agency and free action made in the last section. Specifically, I will defend two claims: first of all, the intuitive judgment triggered by the Frankfurt-Style cases is about free action, rather than free agency. Second, free agency is not necessary for free action. That is, it is possible for an agent to perform a free action even though his free agency is temporally undermined. These two claims lead to the conclusion that the Frankfurt-style cases, if successful, only show that alternative possibilities are unnecessary for free action; it does not show that alternative possibilities are unnecessary for free agency.

### 4.1 Frankfurt-Style Cases are about free action rather than free agency

Since the Frankfurt-Style cases are about moral responsibility, some clarifications are required. Like the dichotomy of free agency and free action, there are two usages in moral responsibility— we can speak of responsible agent and responsible action. When we are speaking of *responsible agent*, the term 'responsibility' refers to a status or an entitlement of the agent. To say that an agent is responsible is just to say that he is entitled to be viewed and treated in a specific way in our moral community.[84] Arguably, this status or entitlement is grounded in a specific kind of free agency, to which I refer as responsible agency. When we are speaking of *responsible action*, the term 'responsibility' refers to a property of the action, or a specific relation that the action bears to its agent.[85] Thus, a similar distinction about moral responsibility judgments can be made:

**Judgment 1**: to judge that the agent is morally responsible for what he has chosen and what he has done.

**Judgment 2**: to judge that the agent possesses responsible agency.

Grant that freedom is the necessary condition for moral responsibility, the distinction between responsible agent and responsible action corresponds to the distinction between free agency and free action. I think

[84] The detail of this entitlement depends on a specific account of moral responsibility. For example, in a Strawsonian account, a responsible agent is entitled to be taken as the target of reactive attitudes. (Strawson 1962)
[85] Similar to last note, I do not fill with the detail so it can be neutral to different accounts of freedom and moral responsibility.

most of the reasoning regarding the Frankfurt-style tend to conflate these two different judgments of moral responsibility.

The Frankfurt-style cases aim to show that alternative possibilities are not necessary for moral responsibility. Since there are two different types of judgments about moral responsibility, we should be cautious to see which type of judgment is made in the Frankfurt-Style case. Obviously, what the Frankfurt-style cases directly invite us to make is the Judgment 1—that is, the judgment about whether the agent in the scenario should be held morally responsible for his action. Now the question is whether the Frankfurt-Style cases also indicate Judgment 2 and show that alternative possibilities are not necessary for responsible agency. My answer is that the Frankfurt-style cases are silent on Judgment 2. When presented a Frankfurt-style case, we are NOT asked whether the agent is a responsible agent. The judgment of responsible agency is suspended in the Frankfurt-Style Cases. Here is an argument for this claim.

In the thought experiment, we are only asked to make the judgment about moral responsibility after the agent has made the decision and acted. However, whether the agent possesses responsible agency seem not to depend on how the agent acted. Let us consider an *incomplete* version of the Frankfurt-Style case:

> Jones is making a hard choice. He can choose either to act in accordance with a moral principle or to act out of his self-interest. Unbeknownst to Jones, a nefarious neuro-scientists, Black has implanted a device into his brain which can monitor his thoughts and manipulate his behaviour. If Jones show any signs indicating he tends to choose to act on the moral principle, Black will interfere and make Jones choose to act on self-interest. Otherwise, Black will let Jones make his own choice.

Now, if we are asked whether Jones possess the relevant moral agency, what would be our response? (A guiding assumption is that before acting and after acting, the agent's agency should be the same.)

The first response is that it depends on how the scenario ends up. Accordingly, we can only know for sure whether the agent possesses responsible agency after we know the result whether he acts on his own. However, this response strikes me as strange since the question of whether the agent is *currently* free should not depend on how he is going to act in the future.

The second response is that the Jones does possess the responsible agency because if it turns out that Jones acts on his own choice and Black does not intervene, then Jones is morally responsible. I think this response is unreliable for two reasons. First, this response speaks against our intuitions. I think most people will feel reluctant to admit that Jones is a responsible agent in this incomplete version of

Frankfurt-style cases. After all, Jones's ability to act on his own choice is severely hindered by the device implanted in his brain. Second, one can equally point out that if the story turns out to end in the other way, then Jones is not responsible for his action. That is to say, even though it is possible for Jones to perform a responsible action, this fact is not sufficient for us to ascribe responsible agency to Jones. For it is equally possible for Jones to trigger the intervention from Black and no moral responsibility is ascribed to the resultant act.

A more reasonable response is that Jones's free agency is undermined to some extent but we do not have the relevant information to fully assess Jones's responsible agency at this point. We only know that Jones's free agency is harmed within a specific choosing situation. However, free agency is a matter of degrees and we are not sure the scope of the influence of Black. Suppose that Black is going to supervise Jones's choice globally such than whenever Jones tends to make a choice that Black is not happy with, then Black will interfere. In such a version of Frankfurt-Style case, we will definitely think that Jones does not possess the responsible agency.[86] Or suppose that Black is going to interfere for one time with respect to Jones's particular choice. In this case, we tend to judge that Jones's responsible agency is temporally and restrictedly undermined but he generally retains it. Nevertheless, the above argument is sufficient to show that in the Frankfurt-style cases, we are making judgment 1 rather than judgment 2. Thus, the Frankfurt-Style cases do not directly show that the ability to do otherwise (or access to alternative possibilities) is unnecessary or irrelevant to responsible agency.

### 4.2 Free agency is not necessary for free action

Some may think that even though the Frankfurt-Style cases do not directly invite us to make Judgment 2, it can still be inferred based on Judgment 1. Specifically, if the fact that the agent possesses free/responsible agency is necessary for the fact that the agent performs a free/responsible action, then from the fact that the agent has performed free/responsible action we can conclude that the agent is a free/responsible agent. In what follows, I will argue that this is not the case. Possessing free agency is not a necessary condition for performing a free action. Sometimes we can make a judgment of responsible action even though free agency is (partly) undermined.

[86] I am not claiming that in this global version of Frankfurt-Style cases, Jones cannot be morally responsible for his action. It may turn out that Jones luckily makes all his choice on his own and never triggers Black's interference so that Jones is morally responsible for every move. Still, we are just ascribing responsibility to every specific action of Jones's from a retrospective perspective. Agency ascription, however, has to be done from a forward-looking perspective.

As shown in the last section, the relation between being free and acting freely is parallel to the relation between possessing an ability to act and performing that action: being free is possessing free agency while acting freely is exercising free agency. Performance and competence are highly correlated. However, performance indicates but does not imply competence. In some situations, performance happens without competence because luck is involved. For example, a novice dart player may hit the bull's eye at his first trial by luck even though he does not possess the ability to reliably hit the bull's eye in a wide range of situations. This suggests the possibility that an agent performs a free action while his free agency is undermined. [87]

Even if the relation between being free and acting freely is a relation of competence and performance, the Dart player analogy is still questionable. We can hardly imagine a person without any piano training can play Chopin. So, maybe acting freely is more like playing Chopin than playing dart. That is, acting freely cannot just be a matter of luck, it has to be explained by the abilities of the agent. Then, how is it possible for Jones to perform a free and responsible action if his free agency is undermined? To answer this question, let's first consider the cases of machines.

> Suppose there is a mechanical glitch with Mike's car such that if Mike steers left, it will turn left; but if he steers right, it will not turn right. Fortunately, when driving home, Mike only steered left in at all the turnings and the car was completely under his control.[88]

Now consider the following questions. Has Mike's car *functioned well* when driving back home? The answer is probably positive. After all, nothing went wrong when the car was working. But if we consider the question of whether Mike's car *is well-functioning*, the answer is probably negative. Well-functioning is a modal property. If a car is well-functioning, it must function well in a broad range of hypothetical scenarios. This is not the case for Mike's car—it would not work properly if it is required to turn left. The lesson is that a factor which harms a machine's competence does not necessarily harm the machine's actual performance. If a machine is just partially mal-functioning such that the glitch only affects a part of the mechanism which is not always operational, then it still functions well as far as the glitch does not show up. Thus, the fact

[87] In the Frankfurt-style cases, Jones manage to perform the free action because of a coincidence that his choice matches Black's plan. The factor of luck becomes significant if we adapt the Frankfurt-style cases into a version of consecutive choice-making such that Jones makes a series of decisions while all of them happen to satisfy Black and does not trigger the intervention.

[88] This case is inspired by Fischer and Ravizza (1998) where they introduce the idea of guidance control.

that the machine is well-functioning is not a necessary condition for the fact that the machine has functioned well.

The above discussion applies to the issues of agency. The idea that the machine is well-functioning is analogous to the idea that the agent is free; the idea that the machine has functioned well is analogous to the idea that the action is free. As shown, a factor which harms a machine's competence does not necessarily harm the machine's actual performance; similarly, a factor which undermines free agency does not necessarily undermine free action. This is exactly the set-up in the Frankfurt-Style cases: the agent is deprived of the freedom to do otherwise by the neuroscientists; while his actual performance is not intervened upon. If free action is possible without free agency, then we are in a position to answer the question in virtue of what Jones performs a responsible action with his free agency being undermined. Recall that we have two ways to tell that the action is free. First, we can tell an action is free if it is performed by a free agent. In this way, the judgement of free action is derived from the judgment of free agency. In another way, we can tell the action is free if it is generated in the right way. That is, we can tell an action is free even if the judgment of free agency is suspended. In the Frankfurt-Style cases, nothing is presumed about what the right process of action should be. However, since the action is performed without intervention, we can at least guarantee that the action is as free as normal.

So far, I have argued that the intuitive judgments triggered by the Frankfurt cases are about free/responsible action rather than free/responsible agency. In addition, we cannot infer that the agent in the Frankfurt-style cases possess responsible agency from the fact that the agent has performed a responsible action. This is because possessing responsible agency is not necessary for performing responsible action. Therefore, compatibilists can, on the one hand, insist that free agency (particularly, reasons-responsiveness) consists of the conditional ability to do otherwise, while endorsing the intuition from the Frankfurt-Style cases, according to which, an agent is morally responsible for his action despite not being able to do otherwise.

If free agency is not necessary for free action, then how should we understand the relation between free agency and free action? As mentioned, I take the relation between free action and free agency as a type of competence-performance relation. Since competence is a reliable indicator of performance, we can also take free agency as a reliable indicator for free action. Specifically, if we know that an agent possesses free agency, we can reliably predict that the agent will perform free action in a wide range of scenarios.[89] Similarly, if we know that an agent is accountable, we can reliably predict that the agent will perform responsible action in a wide range of scenarios. But knowing this doesn't yet settle for us the

[89] This echoes Gilbert Ryle's view (1949/1990) that modal concepts are the license for making inference.

precise role of reasons-responsive agency when it issues in responsible action. That will depend on reasons-responsiveness agency being operative in the 'right' way in the production of action. In the next chapter, the details of this proposal will be articulated. Specifically, I will show how the modal considerations related to free agency help to assess whether a process issuing in an action is a right process or wrong one for a responsible action.

In addition, even though from a *metaphysical* point of view, possessing responsible agency is not necessary for performing responsible action, only in special cases (e.g., the Frankfurt-Style cases) can these two come apart. That is, in cases where the factor impairing responsible agency will be so confined that it does not have any actual impact on the action. These cases are rare and unlikely to arise in actual moral practices. Thus, from a *practical* point of view, we can still regard responsible agency as a necessary condition for responsible action.[90]

**Conclusion**

In this chapter, I have defended the view that reasons-responsiveness consists of the ability to do otherwise. More specifically, reasons-responsiveness is the agent's ability to act differently in response to sufficient reasons in counterfactual scenarios.

In the first part, I have argued against two accounts which dissociate reasons-responsiveness with the ability to do otherwise, namely, Fischer and Ravizza's mechanism-based reasons-responsiveness account and Sartorio's causal reasons-sensitivity account. The problem with Fischer and Ravizz's account is that it faces a dilemma: either their account is not sufficient to ground moral responsibility; or that it involves *ad hoc* postulate and collapses to a leeway approach. The problem with Sartorio's account is that it fails to accommodate two features of agency, which is future-orientation and modality.

In the second part, I have argued that the tension between the leeway compatibilism and source compatibilism can be relieved. This is achieved through a distinction between free/responsible agency and

[90] Given the speciality of the Frankfurt-style cases, some authors suggest we can restrict its relevance to our actual moral practice even we can endorse the basic intuitions from those cases. For example, Horgan (2015) proposes that the Frankfurt-style cases only imply that the access to alternative possibilities is a defeasible necessary condition for moral responsibility. Similarly, Whittle (2016) argues that we should only add a '*ceteris paribus*' clause to the PAP principle in order to accommodate the input from the Frankfurt-style cases. Both Horgan and Whittle maintain that in normal or everyday situations, alternative possibilities are necessary for moral responsibility.

free/responsible action. Specifically, I propose that the leeway compatibilists are led by the question of what grounds free agency, and the source compatibilists are led by the question of what grounds free action. Finally, I have shown that with this distinction, we can, on the one hand, accept the intuitive judgment from the Frankfurt-Style cases that the agent is morally responsible for his action despite not being able to do otherwise; on the other, we can insist that the ability to do otherwise is necessary for free and responsible agency.

However, there are important questions unsettled. First, it is not clear how does the notion of free agency and free action relate to one another. Specifically, if free agency is understood as reasons-responsiveness, then how is reasons-responsiveness relevant to the actual occurrence of action? In addition, in this chapter, I have only shown that my reasons-responsiveness theory is at best compatible with the intuitions triggered by the Frankfurt-Style cases. There is another question of whether my reasons-responsiveness theory is supported by Frankfurt-Style cases. These questions are to be discussed in the next chapter.

## Chapter 3: Reasons-Responsiveness and Explanation

**Abstract:**

In this chapter, I will tackle the challenge of Irrelevance introduced in the last chapter. Specifically, I am going to reconcile a conflict between two theses which are popular among contemporary compatibilists. First, agents' responsible agency is reasons-responsiveness, which is a modal property and analyzed in terms of possible scenarios. Second, responsible agency is grounded in the actual sequence, which is a categorical fact and can be cashed out in terms of causal histories. A key move to reconcile this conflict is to develop an account for the explanatory relevance of reasons-responsiveness.

### Introduction

The reasons-responsiveness theories are one of the most defensible compatibilist theories of free will and moral responsibility. The core idea of these theories is that the ability to recognize reasons and to react to reasons is the defining characteristic of freedom and the necessary control condition for moral responsibility. Different reasons-responsiveness theorists have different accounts for the nature of reasons-responsiveness. In the last chapter, I have argued that reasons-responsiveness consists of the agent's conditional ability to do otherwise. Admittedly, not every proponent of reasons-responsiveness theories holds this view.[91] However, almost all of them hold that reasons-responsiveness is a modal property.[92] That is, reasons-responsiveness is analyzed in modal terms such that an agent A performs a certain action X in a reasons-responsive way only if A would refrain from performing X in certain possible scenarios in which A has sufficient reasons not to X.[93] In this chapter, I refer to this idea as the modal conception of reasons-responsiveness. The main focus of this chapter is to relieve the purported tension between this modal conception of reasons-responsiveness and the Frankfurt-Style cases—a type of thought experiments which have been used to argue against the idea that responsibility requires alternative possibilities.

In the last chapter, I have mentioned that reasons-responsiveness views confront two different challenges from the Frankfurt-Style cases. First of all, reasons-responsiveness, understood as an ability to

[91] E.g., Fischer and Ravizza 1998, McKenna 2015.
[92] The only exception is Sartorio (2015;2016), who develop a de-modalized account of reasons-responsiveness. I have discussed and criticized this view in last chapter.
[93] How to quantify these possible scenarios depend on what degree of reasons-responsiveness is required by moral responsibility.

do otherwise, is incompatible with the direct intuitive judgment from the Frankfurt-Style cases, according to which the agent in the scenario is morally responsible for what he does despite lacking the ability to do otherwise. This intuition implies that the ability to do otherwise is not necessary for moral responsibility. I call this the challenge of Unnecessity. Confronting this challenge, some philosophers try to dissociate reasons-responsiveness with the ability to do otherwise. Notable proposals include the mechanism-based reasons-responsiveness theory from Fischer and Ravizza (1998) and the causal reasons-sensitivity theory from Sartorio (2015; 2016). I have criticized both proposals in the last chapter and argued that the challenge can be resolved by making the distinction between freedom of action and the freedom of agency, and correspondingly, the distinction between responsible action and responsible agency.

Besides, there is a deeper tension between reasons-responsiveness theory and the Frankfurt-Style cases. According to the modal conception, reasons-responsiveness is characterized in terms of counterfactual scenarios. In stark contrast, the Frankfurt-Style cases motivate a new model of freedom, namely, the actual-sequence view, according to which free action depends exclusively on the *actual sequence* leading to the action. The direct intuitive judgment tends to make reasons-responsiveness unnecessary for freedom and moral responsibility; while this actual-sequence view tends to make reasons-responsiveness an irrelevant notion to freedom and moral responsibility. I call this the challenge of Irrelevance. This challenge has recently been flagged up by Sartorio (2015; 2016). This is bad news for the proponents of reasons-responsiveness theories for many of them are either convinced by the Frankfurt cases or employ these cases as support for their theories. If it turns out that the modal conception and the actual-sequence view is inconsistent, it will make the reasons-responsiveness theories internally inconsistent. This chapter is an investigation of the challenge of Irrelevance to the reasons-responsiveness theories. I will reveal that this challenge boils down to specifying the explanatory relation between reasons-responsiveness and the action. To tackle this challenge, I will develop an account for the explanatory relevance of reasons-responsiveness.

The chapter will be structured as follows: In the first section, I will introduce the Frankfurt-Style cases and give a hint about how these cases purported to challenge an influential reasons-responsiveness theory developed by Fischer and Ravizza. In the second section, I will reconstruct Sartorio's argument for the inconsistency between the Frankfurt-Style cases and the reasons-responsiveness theories in detail. In the third section, I will diagnose that the problem is rooted in the explanatory status of reasons-responsiveness. Specifically, as an unmanifested modal property, reasons-responsiveness seem not to be explanatory. Based on this diagnosis, I will provide an improved argument for the inconsistency between the Frankfurt-Style cases and the reasons-responsiveness theories. The key claim is that reasons-responsiveness is irrelevant to moral responsibility if it is explanatorily irrelevant to the explanation of

action. In addition, I will expand the challenge by introducing the explanatory hypothesis of moral responsibility. I will show that if reasons-responsiveness is not explanatory, then reasons-responsiveness theories are not only inconsistent with the Frankfurt-Style cases, but also inconsistent with our everyday moral practice. The real challenge then is to provide an account of how the notion of reasons-responsiveness explains the occurrence of the agent's decision or action. I will take on this challenge in the final section. I will develop an account under Lewis's liberal model of causal explanation. I will argue that reasons-responsiveness is explanatory in virtue of providing information about the causal history of actions. This shows the causal relevance of reasons-responsiveness, and so meets the second key challenge I identified to reasons-responsiveness accounts. In the concluding section, I will briefly envision a strategy about how the compatibilists can fight back against the incompatibilists on the foothold of the Frankfurt-Style cases.

**1. The Frankfurt-Style Cases and Reasons-Responsiveness**

There was a time when the most central issue in free will debates was whether the ability to do otherwise is compatible with determinism. Now, attention has considerably shifted from the notion of ability to do otherwise to the notion of moral responsibility. Quite a few contemporary compatibilists concede that determinism is not compatible with the ability to do otherwise, while they insist that determinism is compatible with the freedom or the control required by moral responsibility.[94] This change is mainly due to the impact of the Frankfurt-style cases, a kind of thought experiments which attack the assumption that moral responsibility requires the ability to do otherwise. To be more precise, these thought experiments try to challenge a widely held principle called the Principle of Alternative Possibilities (PAP):

> PAP: An agent is morally responsible for performing a given act A only if he or she could have done otherwise.[95]

This kind of thought experiments were first put forward by Harry Frankfurt (1969). A typical Frankfurt-Style Case is a scenario which features a counterfactual intervener who has a disposition to make the agent perform a certain action only if the agent has a tendency to refrain from doing that action. Such a stipulation robs the agent's ability to do otherwise. In addition, the scenario ends up with the fact that the

[94] E,g., Fischer and Ravizza (1998), McKenna (2013), Sartorio (2016).
[95] See Frankfurt (1969, 829). Frankfurt's original term for this principle is 'the principle of alternate possibilities'.

agent chooses to do that action on her own reasons, which means that the counterfactual intervener does not intervene. These thought experiments aim to show that alternative possibilities are not necessary to attribute moral responsibility to the agent. Recall the typical Frankfurt case we met in the last chapter:

> Jones is an agent confronting a hard choice whether he will cheat on the exam. Unbeknownst to Jones, a resourceful neuroscientist, Black wants to ensure that Jones cheats on the exam. Black has implanted a tiny chip into Jones brain when Jones was asleep. This chip can not only enable Black to monitor Jones's brain activities but also makes Jones decide in accordance with Black's will. Black is going to make Jones decide to cheat only if he discovers that Jones is inclined not to cheat. Otherwise, Black will let Jones to make his own decision. It turns out that Jones decides to cheat on his own.

Many compatibilists share the intuition that the agent (Jones) is morally responsible for his action even though he lacked access to alternative possibilities other than cheating on the exam. Therefore, they conclude that moral responsibility does not require alternative possibilities or the ability to do otherwise. Since it is widely assumed that a relevant kind of freedom is required by moral responsibility, the Frankfurt-Style cases are also used to favour the view that the relevant freedom for moral responsibility does not require the ability to do otherwise.

Motivated by the Frankfurt-Style cases, many compatibilists subscribe to a new model of free agency—the *actual-sequence view*. Contra the more traditional alternative-possibilities views which hold the access to the alternative possibilities is the key to freedom, the actual-sequence view can be summarized as the idea that whether the action is free depends on the actual-sequence leading to that action. This model is friendly to compatibilism.[96] Within this model, acting freely is no longer a matter of whether the action is determined or not, but a matter about whether the action is determined *in the right way*. Thus, an action can be free even in a deterministic world if it comes about through the right actual sequence.

But what is the right actual sequence? One suggestion is that it should be a causal history of action without control-undermining factors such as manipulation, coercion, duress, hypnotism, irresistible urge, psychological disorders and so on.[97] This list of control-undermining factors can extend infinitely. Instead of adding the potential control undermining factors in a somewhat *ad hoc* way, it would be better if we could give an account of what features the agent has to have—what positive conditions there are for

[96] As I have shown in the last chapter, these two views are not competing, but complementary. They aim to answers different questions—what grounds freedom of agent and what grounds freedom of action.
[97] E,g., Fischer and Ravizza (1998), McKenna (2013), Sartorio (2016).

responsibility (which are undermined by the presence of those factors). Here are how the reasons-responsiveness theories come into play. Many compatibilists try to delineate the specific kind of freedom or control condition for moral responsibility in terms of reasons-responsiveness. That is, when the agent is reasons-responsive, she is free from those freedom-undermining factors in acting.

The most influential and sophisticated reasons-responsiveness theory is defended by Fischer and Ravizza's (1998), which I have discussed in the last chapter. On the one hand, Fischer and Ravizza believe that freedom and moral responsibility is grounded in reasons-responsiveness. On the other hand, they are convinced by the Frankfurt-Style cases and contend that the ability to do otherwise is not necessary for freedom and moral responsibility. As mentioned, these two commitments lead to an apparent conflict—if reasons-responsiveness consists of the ability to do otherwise (specified as that 'the agent would do otherwise were there sufficient reason to do so'), then how does it fit with the implication from the Frankfurt-Style cases which show that the ability to do otherwise is unnecessary for freedom? This is the Challenge of Unnecessity which I introduced in the last chapter. To handle this problem, Fischer and Ravizza have to make some modification on the notion of reasons-responsiveness. Instead of locating the free agency on the agent's reasons-responsiveness, they suggest focusing on the mechanism which issues in the action (say, the mechanism of deliberation or practical reasoning and the related motivational system). They argue that in ascribing moral responsibility, what at issue is not the agent's being reasons-responsive, but the mechanism's being reasons-responsive. Roughly, the mechanism is reasons-responsive if and only if, if there are some counterfactual scenarios in which that mechanism operates, the agent would recognize and react to the sufficient reasons.[98] According to their account, the agent's responsibility and freedom are grounded in the mechanism-based reasons-responsiveness, which is intact even when the counterfactual intervener is present in the Frankfurt-Style scenario. This is how Fischer and Ravizza's account solve the challenge of Unnecessity.[99]

Even if the mechanism-based account can handle the challenge of Unnecessity, it confronts a deeper challenge from the Frankfurt-Style cases. The notion of mechanism-based reasons-responsiveness is a modal property which is analyzed in terms of counterfactual possibilities. Call it the modal conception of reasons-responsiveness. This conception has caused some worries. In particular, by making the modal property the focus of responsibility, it seems to shift our attention in assessing agents' moral

[98] Fischer and Ravizza think that to adequately ground moral responsibility, the mechanism should be moderate reasons-responsiveness, which consists of *regular reasons-receptivity* and *weak reasons-reacttivity*. Specifically, to be regular reasons-receptive, the mechanism must enable the agent to recognize multiple reasons such that those reasons constitute an understandable pattern; to be weak reasons-reactive, there must exist one counterfactual scenarios in which the mechanism enables the agent to react to the sufficient reason.

[99] Actually, I am not satisfied with this mechanism-based solution. I have criticized it in last chapter.

responsibility from what they *have actually done* to what they *might possibly do* or *might possibly have done*.[100] Why should we care more about an agent's behaviour in possible scenarios than what has happened in the actual scenario, say her in fact failure to respond to the particular reason or her bad quality of will that lead to her act in the actual scenario?[101] Recently this worry has been raised and articulated in a more sophisticated way by Sartorio. She points out that reasons-responsiveness, understood as a modal property, seem not to be part of the actual sequence; while according to the actual-sequence view, the agent's freedom and responsibility is just a matter about how the actual sequence of the action unfolds. This means that reasons-responsiveness—understood as how the agent would respond were there other sufficient reasons—is irrelevant to agent's freedom and moral responsibility. Thus, there is a tension between the modal conception of reasons-responsiveness and the Frankfurt-Style cases which motivate the actual-sequence view. If this argument works, it is bad news for Fischer and Ravizza. They heavily rely on the Frankfurt-Style cases to tackle the incompatibilist challenges such as the consequence argument (as we saw in the last chapter). They also subscribe to the actual-sequence view. Specifically, their mechanism-based reasons-responsiveness theory is the result of adapting the reasons-responsiveness into an actual-sequence framework. Thus, the tension between the modal conception of reasons-responsiveness and the actual-sequence view indicates that their position is internally inconsistent. Note that this objection would also face the novel account of reasons-responsiveness I advanced in the last chapter, which holds that reasons-responsiveness consists of the conditional ability to do otherwise. My account faces the same worry because the conditional ability to do otherwise is a modal property which is cashed out in counterfactual conditionals.

## 2. Sartorio's Argument for the Purported Inconsistency

In this section, I will lay out Sartorio's (2016) argument for the claim that the modal view of reasons-responsiveness is incompatible with the actual-sequence view.[102] There are two steps in her arguments: first, she argues that the Frankfurt-Style cases support the actual-sequence view, which can be further interpreted as a supervenience claim about freedom. According to this claim, freedom supervenes on the actual causal history. Second, Sartorio tries to establish an inconsistency between the supervenience claim

[100] For this point, see McKenna (2005).
[101] For example, Strawson (1962) famously contends that people's reactive attitudes, which are the hallmarks of moral responsibility ascription, track the quality of will of whom are being praised or blamed.
[102] Sartorio uses the term 'reasons-sensitivity' to mean reasons-responsiveness. To keep the terminology consistent, I will keep using 'reasons-responsiveness'.

and the modal view of the reasons-responsiveness theory by introducing the case of Frank and Insenstive Frank. In this hypothetical scenario, two agents have similar actual causal histories while differing in reasons-responsiveness. I set out each argument in turn below.

### 2.1. Argument for the Supervenience Claim

Like many other compatibilists, Sartorio agrees that the Frankfurt cases have convincingly shown that the alternative possibilities are irrelevant to agents' freedom. In particular, Sartorio extracts two relevant intuitions from the Frankfurt-Style cases:

> Intuition 1: The agent (in a Frankfurt case) is in control of his act despite his lack of robust alternatives.
>
> Intuition 2: What determines whether the agent is in control of his act is how he actually came to perform the act. [103]

Both of these two intuitions are against the traditional alternative-possibilities view of free agency (or sometimes called the leeway freedom). Intuition 1 claims that the alternative-possibilities are not necessary while Intuition 2 claims that the relevant factors are the events that occurred in the *actual* sequences. In addition, Sartorio contends that Intuition 2, which buttresses Intuition 1, helps to motivate the actual-sequence view of freedom.[104] The actual-sequence view can be further divided into two claims—the positive claim and the negative claim.

> The positive claim (P): the freedom of agents is grounded, at least partly, in actual sequences (and the ground of actual sequences).[105]

[103] Sartorio 2016, 17
[104] Sartorio 2016, 18
[105] Sartorio is open to the idea that the actual sequence or the actual causal history can be further grounded in modal facts. Sartorio (2018 a; 2018 b) gives the example that causation under certain accounts can be further grounded in modal facts. Thus, it is possible that freedom is grounded in the causal history, while the causal history is further grounded in modal facts. However, even in these versions of actual-sequence view, freedom is not directly grounded in modal facts.

> The negative claim (N): The freedom of agents is not grounded in anything other than actual sequences (and the grounds of actual sequences).[106]

These two claims lead to the idea that freedom is grounded *exclusively* in the actual sequences. Note that both claims involve the metaphysical notions of *grounding*. Sartorio proposes that investigation on the nature of free agency is about looking for the grounding conditions for freedom. This departs from the more traditional enterprise of just looking for the sufficient and necessary conditions for freedom. Roughly, to say X grounds Y is just to say Y obtains in virtue of X. In other words, to study the nature of free agency, we should find out in virtue of what the agent is free.

What the positive claim says is that the agents act out of freedom (at least partly) in virtue of the actual sequence which issues in the actions. Using the Frankfurt case presented above as an example, the positive claim tells us what are relevant to Jones's action are those mental antecedents that lead to Frank's action in the actual sequence, say his motives, desires, decisions and intentions. If there are some freedom-undermining factors in the actual sequence such as coercion, physical constraints or phobia, then the agent's freedom will be compromised or entirely deprived. In addition, since the actual sequences may be grounded in some further conditions, so agents' freedom may be grounded in those further conditions which ground the actual sequences. By contrast, the negative claim tells us that factors apart from actual sequences are irrelevant to agents' freedom. Although at this stage the negative claim is a little vague since it is not clear what actual sequences are, this claim specifically denies the idea that freedom of agents is grounded in alternative possibilities or the ability to do otherwise. Sartorio contends that both of these claims are necessary for a complete articulation of the actual-sequence view: The negative claim is to differentiate the actual-sequence view from the possible-alternatives view; the positive claim is to make clear that freedom is grounded in something rather than nothing. After presenting the outline of the actual-sequence view, Sartorio suggests that the best interpretation of the actual sequence views is that freedom is grounded in actual causal facts— "facts that certain events are connected in certain ways".[107] This interpretation makes it clear that the a*ctual sequences* in Sartorio's characterization of the actual-sequence view are *causal histories*.

She then argues that such an interpretation of the actual-sequence view implies a supervenience claim about freedom-- an agent's freedom with respect to a certain action supervenes exclusively on the actual causal histories issuing in that action. In other words, an agent's freedom is a function of their causal

[106] Sartorio 2016, 18
[107] Sartorio 2016, 21

histories. The variables which enhance or reduce an agent's freedom must be elements of their causal histories. Sartorio expresses this idea in a slogan: *No difference in freedom without a difference in the causal sequence*.[108] For example, if two agents A and B perform similar actions but A is free and morally responsible while B is not, then a difference must be found in the actual sequences leading to their actions — say, B is forced to perform the action while A acts on her own.

Given that the causes in the causal histories are abundant and that not all the causes in the causal histories are relevant to agents' freedom, Sartorio suggests that supervenience claim should be qualified. That is, freedom supervenes on the relevant part of the causal history rather than the whole causal history. She calls it the Strong Supervenience Claim.

> The Strong Supervenience Claim: An agent's freedom with respect to X supervenes on those elements of the causal sequence issuing in X that ground the agent's freedom.[109]

Up till now, Sartorio has developed her supervenience claim for freedom. Here is a summary of her argument. First, Sartorio identifies two intuitions about the Frankfurt cases—Intuition 1 and Intuition 2. Both intuitions are against the alternative-possibilities view of freedom and Intuition 2, in particular, motivates a competing view of freedom, namely, the actual-sequence view of freedom. According to the actual-sequence view, agents' freedom is grounded in exclusively on the actual sequence (and the grounds of actual sequences). Sartorio further interprets the actual sequence as the causal history of the action. This implies a supervenience claim about freedom, according to which agents' freedom supervenes on the relevant elements of the causal histories.

### 2.2 Argument for the Inconsistency: Frank and Insensitive Frank

The tension between the actual-sequence view and the modal conception of reasons-responsiveness can be outlined in the following way: on the one hand, according to the actual-sequence view and the derived supervenience claim, an agent's freedom with respect to an event is exclusively grounded in the actual causal histories for the action. However, reasons-responsiveness as a modal property is hard to fit into the causal history. To highlight this tension, Sartorio make uses of the supervenience claim demonstrated above. According to the supervenience claim, there is no difference in freedom and moral responsibility without difference in the causal histories. If it is possible that there are two agents with different reasons-responsiveness while they have the same causal histories for actions, then it will imply that the

[108] Sartorio 2016, 29
[109] Satorio 2016, 29

supervenience claim is violated by the modal conception of reasons-responsiveness. She devises a pair of scenarios feature with two different agents—Frank and Insensitive Frank. They perform similar actions. Frank shoots his enemy Furt; and insensitive Frank shoots the counterpart, say Furt*. In addition, they have the sane actual causal histories of action in relevant respects. We can assume that Frank's action is driven for a desire to revenge; and so is Insensitive Frank's. The only difference between Frank and insensitive Frank is that Frank can respond to reasons while insensitive Frank cannot. For example, when Frank learns that Furt is the father of five children who depend on him to survive, this will constitute a sufficient reason for Frank to refrain from killing. By comparison, even when Insensitive Frank learns that Furt* is the father of five children who depend on him to survive, Insensitve Frank will not change his mind.

> …unlike Frank, Insensitive Frank is insensitive to most kinds of reasons to refrain from shooting [Furt*] (and not, let's assume, because of some earlier free decision to become insensitive to reasons in this way): he is such that, in most types of counterfactual scenario where those reasons are present, he is not receptive to those reasons or is not sufficiently motivated by them, and thus his desire for revenge still prevails.[110]

Given that both Frank and Insensitive Frank killed their enemies through same causal histories in relevant respects while they are different in reasons-responsiveness, the reasons-responsiveness theory and the actual-sequence view indicate different results of freedom and responsibility attribution. According to the reasons-responsiveness theory, Frank was free with respect to his action while insensitive Frank was not because insensitive Frank did not act in a reasons-responsive way.[111] However, according to the actual-sequence view, Frank and insensitive Frank should be equally free and equally responsible with respect to their actions because the causal histories issuing in their actions are identical to each other.

Sartorio challenges the reasons-responsiveness theory by asking the question of how the difference of reasons-responsiveness between Frank and insensitive Frank can be 'reflected' in the causal histories. She argues that according to the reasons-responsiveness theory, the differences between Frank and Insensitive Frank are "purely counterfactual".[112] That is to say, the reasons-responsiveness theorists can only explain the differences by citing counterfactual scenarios: if there were sufficient reasons then Frank would

[110] Sartorio 2016, 119–20.
[111] It is dubious that insensitive Frank was totally unfree or irresponsible with respect to his action. From the description of the scenario, insensitive Frank decided to kill Furt at least for some specific reasons. If those reasons did not obtain, it would be reasonable to assume that Furt would not decide to kill Furt. It might be suggested that insensitive Frank is just an extremely persistent and tenacious person who will not give up his plan. But he still acts on reason thus he is responsible for his action in some sense. My intuition is, the situation of insensitive Frank is quite different from the situation of an insensitive addict, whose behaviour is driven by irresistible urge or desire.
[112] Sartorio 2016, 120.

refrain from killing while insensitive Frank would not. But this move may violate the actual-sequence view because it accounts for freedom of agents with reference to factors beyond actual sequences.

Sartorio entertains two possible responses and argues against them respectively. The first is that reasons-responsiveness can be reflected in the strength of the desire. She thinks that this proposal falsely presupposes that resistibility of desire is just a function of strength.

The second response is that the difference of reasons-responsiveness is reflected in the underlying physical constitutions. It is plausible that the neural structures in the brains of Frank and Insensitive Frank are different, which account for the differences of reasons-responsiveness between the two agents. Accordingly, the causal histories issuing in Frank's action is different from the causal histories issuing in insensitive Frank's action because their actions were issued in through different mechanisms with different physical constitutions. However, Sartorio refutes this solution. She argues that "in identifying the actual mechanism issuing in a choice we must look only at the elements that are in fact causally operational".[113] Thus, even if Frank's physical constitutions are different from insensitive Frank's, it is possible that the causally operational parts remain the same; and the different parts would only be operational in the possible scenarios where Frank refrained from acting with sufficient reasons.

Sartorio then concludes that the modal view of reasons-responsiveness is inconsistent with the supervenience claim. In consequence, the modal view of reasons-responsiveness is not consistent with the actual sequence view of freedom either. Specifically, if reasons-responsiveness does not pertain to the actual sequence, then according to the actual-sequence view, reasons-responsiveness is not relevant to moral responsibility. This is the challenge of Irrelevance which I have introduced in the last chapter. This challenge particularly confronts those compatibilists who rely on the intuitive resources from the Frankfurt-Style cases and try to embed the reasons-responsiveness theory into the actual-sequence view.[114] There are two routes to avoid this inconsistency: either compatibilists need to abandon the modal conception of reasons-responsiveness to account for free agency, or they need to abandon the actual-sequence view and the Frankfurt-Style cases. However, both routes seem not to be palatable. Abandoning the intuitions about the actual-sequence view generated by the Frankfurt-Style cases means to abandon a powerful intuitive resource to tackle the incompatibilist challenge; while abandoning the modal conception means to abandon a fundamental consideration of free agency.[115] I do think there is a solution

[113] Sartorio 2016, 121
[114] Typically, Fischer and Ravizza (1998); Fischer (2012); McKenna (2013)
[115] Actually, this is what Sartorio set out to do. She develops the causal reasons-sensitivity account, according to which the agent's freedom (or reasons-sensitivity) is exclusively grounded in the actual causal history. The trick is to allow the causal actual history comprise not only reasons but also absences of reasons. which a de-modalized view of reasons-responsiveness. I do not think this is a successful solution and I have criticized it in the last chapter.

to this problem. Before presenting my solution and integrating it into my account of free agency that I outlined in the last chapter, I will first make the challenge more compelling and general.

### 3. Reasons-Responsiveness and Explanation

An important move of Sartorio's argument is to show that it is *intuitively* possible for two agents to have similar causal histories in actions while being different in reasons-responsiveness (the case of Frank and Insensitive Frank). This subsequently implies that reasons-responsiveness is not part of the actual sequence. However, merely relying on intuition is not adequate to establish this conclusion. Other reasons-responsiveness theorists may simply deny this intuition and insist that reasons-responsiveness does pertain to the actual sequence.[116] This controversy is not to be resolved if it is still not clear what the actual sequence is, and by what principle we can tell a fact pertains to the actual sequence or not.

In this section, I want to entertain some independent motivations for the claim that reasons-responsiveness is not part of the actual-sequence. Specifically, I argue that that reasons-responsiveness is ostensibly not part of the actual sequence because it is ostensibly not explanatory. And reasons-responsiveness is ostensibly not explanatory because it is taken as an unmanifested disposition or an unexercised ability. Based on these diagnoses, I will provide a new argument for the inconsistency between the modal conception of reasons-responsiveness and the Frankfurt-Style case. Besides, once it is made explicit that the nature of the problem is rooted in the explanatory status of reasons-responsiveness, then the scope of the challenge will be expanded. That is, the modal conception of reasons-responsiveness is not only inconsistent with the rationale motivated by Frankfurt-Style cases, it is also inconsistent with our everyday responsibility practice.

#### 3.1 Reasons-Responsiveness and Actual Sequence

Are there any arguments to rule out reasons-responsiveness from being part of the actual sequence leading to the action? The first suggestion is that the Frankfurt-Style case implies that reasons-responsiveness does not pertain to the actual sequence. To evaluate this suggestion, we should make it

[116] For example, Fischer and Ravizza commit to a more liberal actual-sequence view. They do not clarify what actual sequence consist of. Rather, they just emphasize that an actual sequence which leads to free action should exclude specific control-undermining factors. McKenna (2013) suggests that a modal property can be part of the actual sequence for it tells us something about the actual operation. Thus, to establish the claim of inconsistency, there should be an argument for the claim that being reasons-responsiveness is not a component of the actual causal history.

clear the relation between the Frankfurt-Style cases and the actual-sequence view. Recall that Sartorio's argument begins with two different intuitions about the Frankfurt case.

> Intuition 1: The agent (in a Frankfurt case) is in control of his act despite his lack of robust alternatives.
>
> Intuition 2: What determines whether the agent is in control of his act is how he actually came to perform the act.

Intuition 1 is directly triggered by the Frankfurt-Style cases and it is widely shared by many philosophers. This direct intuition implies that the ability to do otherwise is not necessary for moral responsibility. Intuition 2 is stronger than Intuition 1 for it implies that access to alternative possibilities is not only unnecessary but also irrelevant to moral responsibility.[117] Rather than being directly triggered by the Frankfurt-Style cases, Intuition 2 is more like a theoretical postulate to explain intuition 1. Thus, one can accept Intuition 1 without endorsing Intuition 2. Consequently, since intuition 2 is the basis for the actual-sequence view, one can accept Intuition 1 while rejecting the actual-sequence view. For example, some source incompatibilists may argue that the agent in the scenario is morally responsible for his action only because his formation of the character can be traced back to a certain source in the history over which he has alternative possibilities control.[118] Even if we grant that the actual-sequence view can still be taken as a plausible *compatibilist* rationale to explain the direct intuition triggered by the Frankfurt-Style cases, the Frankfurt-Style cases do not have any indications on the constitution of actual sequence, let alone telling us that reasons-responsiveness is not part of it. Thus, it is left to philosophers' discretion to interpret what an actual sequence is constituted by. [119]

The second suggestion is that actual sequence by definition is constituted only by actual facts. Specifically, if an actual sequence is understood as the relevant part of a causal history leading to the action, it should be constituted by the relevant elements of the causal histories. Since the agent's being reasons-responsive is a counterfactual fact, it is then not part of the actual sequence. But what is an actual fact or a counterfactual fact? [120] The distinction may be made from a semantic point of view. Actual facts are those expressed by true unconditional propositions. By comparison, counterfactual facts are facts cashed out by counterfactual conditionals. Since an agent's reasons-responsiveness is cashed out in counterfactual conditionals, it is a counterfactual fact and does not pertain to the actual sequence. But this

[117] For this distinction, see Chapter 6.
[118] E.g., Kane 1996.
[119] See note 116 above.
[120] The dichotomy of 'actual fact' and 'counterfactual fact' are introduced by Sartorio (2016), but she does not clarify what these two notions mean.

conception of actual sequences is too narrow. There are many facts which are characterized by counterfactual conditionals and are intuitively relevant to the occurrence of events. According to some popular theories, the laws of nature or dispositional properties of objects are analyzed with conditionals. Both the laws of nature and dispositional properties can be relevant to account for events occurring in the actual sequences.[121] For example, the Newton's first law of motion can be characterized as: in an inertial reference frame, if an object was not acted upon by a force, it would either remain at rest or continues to move at a constant velocity, unless acted upon by a force. This law is characterized in counterfactual terms, but it by no means suggests that this law has nothing to do with the events that occurred in the actual causal history. More importantly, even Sartorio herself allows the possibility that counterfactual facts pertain to the actual sequence. According to her specification, there are two ways for a fact to be relevant to an actual sequence—either that the fact is a component of the actual sequence or that the fact helps to ground the actual sequence. In addition, if the fact helps to ground the actual sequence, the fact can either directly ground the actual sequence or serve as a further ground for those that directly ground the actual sequence. Sartorio admits that the actual sequence can be possibly grounded (indirectly) in counterfactual facts. This is because causation may be analyzed in a counterfactual way.[122] This, of course, makes it even more difficult to clarify what an actual sequence is and to explain why reasons-responsiveness is not part of it.

The final and the most plausible suggestion is that elements of the actual sequences should be explanatorily relevant to the events in the actual sequences.[123] Since reasons-responsiveness is irrelevant to the causal explanation of the action, it is not part of the actual sequence. But why is reasons-responsiveness explanatorily irrelevant to the action? The answer should not be merely that reasons-responsiveness is taken as a modal property such as dispositions or abilities for dispositions and abilities can be explanatory. Rather, the answer should be that reasons-responsiveness is understood as a *unmanifested* disposition or an *unexercised* ability. Intuitively, an unmanifested disposition or an unexercised ability is not explanatory. Suppose that a fragile glass falls onto the ground but it does not get broken. Intuitively, the fragility is explanatorily irrelevant to the fact that the glass is not broken.[124] Likewise, reasons-responsiveness is not only understood as a modal property, but a modal property which

[121] For the explanatory relevance of dispositions, see Mckitrick (2005).
[122] See Sartorio (2018a; 2018b). Sartorio also holds that the contrast between actual-sequence view and the alternative-possibilities view are not logical contraries. See Sartorio (2016, 11-12).
[123] Sartorio (2016) in her book does not make a rigid distinction between causation and causal explanation when she talks about actual sequences and actual causal histories. In a symposium paper, Pereboom (2018)suggests to focus on causal explanation rather than causation for the former notion triggers less metaphysical controversies. In her reply, Sartorio (2018a) accepts this suggestion.
[124] This requires more detailed specification. In the next section, I will show that an unmanifested modal property is not explanatory only under a certain model of causal explanation.

is not manifested in the actual history.[125] Recall the case of Frank. Frank is reasons-responsive in shooting Furt in the sense that Frank would *refrain* from shooting Furt for certain reasons. This unmanifested property seems not to figure in the explanation for the actual outcome that Frank shoots Furt.[126] Under this interpretation of actual sequence, the argument is on the right track. If this real issue is about the explanatory status of reasons-responsiveness, then we can disregard the idea of the actual-sequence view and directly establish the inconsistency between the modal conception and the Frankfurt-Style cases.

### 3.2 An Improved Argument for the Inconsistency

I have investigated the several motivations for ruling out reasons-responsiveness from being part of the actual sequence. There are two take-home messages. First, the notion of actual sequence is not well defined. It is not clear what an actual sequence consists of (e.g., events? dispositions? laws of nature?). [127] Because of this lacuna, it is not obvious that reasons-responsiveness does not pertain to the actual sequence which leads to the action. Second, the real issue lies in the explanatory status of reasons-responsiveness. Specifically, reasons-responsiveness is a unmanifested modal property and *ipso facto* not explanatory to the action. Based on these two points, I now propose a more straightforward argument to establish the inconsistency between the reasons-responsiveness view and the Frankfurt-Style case; an argument that highlights the role of explanation.

Recall that the actual-sequence view is postulated to explain the intuition triggered by the Frankfurt-Style cases. The direct intuition is that the agent is morally responsible despite the fact that his ability to do otherwise is robbed by the counterfactual intervener. To explain this intuition, philosophers postulate the actual-sequence view, according to which, what matters to the agent's acting freely is the actual sequence of the action. Since the counterfactual intervener has no impact on the actual sequence, the agent is free. As argued, the notion of actual sequence is not clear enough.

---

[125] Whether reasons-responsiveness is understood as an ability unexercised or a disposition unmanifested depends on whether an agent-based approach or a mechanism-based approach is taken. For this point, see my last chapter.

[126] Admittedly, some may argue that reasons-responsiveness is not a purely counterfactual property. Rather, it composed of a counterfactual part and an actual part: An agent is reasons-responsive in φing if he can refrain from φing for potential reasons; in addition, when the agent is φing for an actual reason, he is also exercising his reasons-responsiveness. That is to say, whenever the agent is acting for a reason, he is exercising his reasons-responsiveness. I am open to this suggestion. However, reasons-responsiveness theorists will insist that it is the counterfactual part or more precisely, the unmanifested part of reasons-responsiveness that helps to ground moral responsibility. Thanks to James Lennman who helps me to see this point.

[127] For a similar concern, see Whittle (2018, 63–64). Whittle argues that Sartorio's distinction between the actual-sequence view and the alternative-possibilities view is not significant because the notion of actual sequence is unclear. For a reply, see Sartorio (2018b)

Fortunately, there is a more straightforward to explain the direct intuition from the Frankfurt-Style cases. That is, since the presence of the counterfactual intervener is irrelevant to the explanation of the agent's action, it is irrelevant to the agent's freedom or moral responsibility. Actually, this is what Frankfurt seems to endorse. In his seminal article, he writes, "when a fact is in this way irrelevant to the problem of accounting for a person's action it seems quite gratuitous to assign it any weight in the assessment of his moral responsibility."[128] This explanation involves less commitment than the actual-sequence view for it only claims what is irrelevant to agents' freedom and moral responsibility while keeping silent on what is relevant. Fischer agrees with Frankfurt on this point and in a recent article he introduces the *Irrelevance Principe* (IP) which articulate Frankfurt's idea more precisely:

> (IP) If a fact is irrelevant to a correct account of the causal explanation of the person's action, then this fact is irrelevant to the issue of the person's moral responsibility. [129] (Fischer 2015, 122)

If IP is taken as the most plausible principle to explain the intuition pumped by the Frankfurt cases, then there is a more straightforward way to demonstrate the tension between the reasons-responsiveness theories and the Frankfurt-Style cases. Specifically, it is difficult to see how the following three statements can be compatible.

> (S1) According to the reasons-responsiveness theories, being reasons-responsive is essential to being morally responsible for action.
>
> (S2) Reasons-responsiveness, conceived as an unmanifested modal property, is irrelevant to the causal explanation for the action.
>
> (S3) The rationale for Frankfurt-Style cases is the IP, according to which if a fact is irrelevant to a correct account of the causal explanation of the person's action, then this fact is irrelevant to the issue of the person's moral responsibility.

It seems that S1, S2 and S3 are all plausible while they cannot be put together (the crucial inconsistency is between S1 and S2). So here is another way to establish the inconsistency between the Frankfurt-Style cases and modal view of reasons-responsiveness: reasons-responsiveness conceived as an unmanifested modal property is not explanatory. There seems to be a close connection between explanation and responsibility. The connection is shown by the Frankfurt-Style cases and captured by the Irrelevance

[128] Frankfurt (1969).
[129] According to Fischer, this principle was first presented by Palmer in an unpublished manuscript. Fischer argues that this principle requires more nuanced qualification to be true. For the sake of simplicity, I will not get into the details and just take for granted that IP is true.

Principle. This argument does not involve the unclarified notion of actual sequence. In addition, it re-directs our attention to the explanatory status of reasons-responsiveness, which I think is the real issue. More importantly, when the focus is on explanation, the challenge of Irrelevance can be made even compelling and general.

### 3.3 Expanding the Challenge: Explanation and Responsibility Ascription

So far, I have presented a new argument for the challenge of Irrelevance, which focuses on the explanatory status of reasons-responsiveness. A crucial premise of the argument is this: reasons-responsiveness is not explanatory (S2). This premise causes trouble to the reasons-responsiveness theories because of the Irrelevance Principle, which is motivated by the Frankfurt-Style cases. Note that this argument does not directly challenge the modal conception of reasons-responsiveness. Rather, it only establishes a tension between the modal conception and the Frankfurt-Style cases. Thus, the challenge restricted *only* to reasons-responsiveness theorists who endorse the intuitions of Frankfurt-Style case at the very beginning. What about those who do not? [130] In what follows, I want to expand the challenge and argue that this challenge is more general and pressing. I will show that Frankfurt-Style cases are not the only route to reach the connection between responsibility and explanation.

Recently, Björnsson and Persson (2012) propose a model of responsibility judgment, according to which responsibility judgment is a kind of explanatory judgment. They call it the *explanatory hypothesis*. The core of this hypothesis is that when we are ascribing moral responsibility to the agent in question, we are making a judgment about whether the agent's motivational structure plays a remarkable role in explaining his action. Accordingly, factors increasing or decreasing the explanatory significance of the motivational structure will influence people's attribution of responsibility accordingly. This model is promising for it both accommodates the phenomena of ordinary moral practice as well as our intuitive reactions to certain philosophical arguments. Specifically, this model accounts for why certain excuses affect responsibility attribution: effective excuses serve as independent explanatory factors which make the motivational structure less or not explanatory. Examples include but not restricted to 'I am out of control', 'I don't know it', 'I was forced', etc. In addition, this model explains our reactive intuitions to certain skeptical arguments against moral responsibility. For example, according to Galen Strawson's basic argument, no one is ultimately morally responsible for his action. This is because, one's action

[130] What I have in mind is Nelkin (2011) and Vihvelin (2004; 2013). Both philosophers deny that the Frankfurt-Style cases successfully show that the agent's relevant ability to do otherwise is robbed in the Frankfurt-Style case while both think that a certain kind of rational ability is the grounding condition for moral responsibility.

results from the *way one is* (e.g., character, values, preferences,) while *the way one is* comes from remote factors such as heredity, early experience and environmental influences, over which one has no control.[131] The model explains why we find such argument compelling—this argument shifts our attention from the agent's motivational structure to more abstract explanatory factors such as heredity and early experience.

If we grant this model of responsibility judgment, we have a way independent of the Frankfurt-Style cases to establish the connection between responsibility and explanation and then question the relevance of reasons-responsiveness to moral responsibility. For if reasons-responsiveness is not explanatory, then *a fortiori* it will not have any impacts on the explanatory significance of the agent's motivational structure. The implication is that the challenge of Irrelevance to reasons-responsiveness becomes more general and pressing: whether reasons-responsiveness theorists endorse the intuitions from the Frankfurt-Style cases or not, they may still confront this challenge (if they endorse the (very plausible) explanatory hypothesis of moral responsibility).[132]

## 4. Defending the Modal View of Reasons-Responsiveness

In this section, I will develop an account for the explanatory status of reasons-responsiveness. This consists of two steps. In the first step, I will demonstrate that reasons-responsiveness is not explanatory because a certain model of causal explanation is postulated. I call it the strict model of causal explanation. I will point out some limitations of this model, which constitutes the motivation to introduce David Lewis's liberal model of causal explanation. In the second step, I will develop my account based on Lewis's model.

[131] See Galen Strawson (1994).

[132] At this point, careful readers may wonder how this challenge of Irrelevance relates to the ideas discussed in the last chapter. In the last chapter, I have made a distinction between responsible agency and responsible action. I have argued that reasons-responsiveness, as responsible agency, is not *necessary* for performing a responsible action (e.g., in the Frankfurt-Style cases). This chapter concerns a different question—whether reasons-responsiveness is *relevant* to responsible action. Particularly, the claim that reasons-responsiveness is not necessary for responsible action is weaker than the claim that reasons-responsiveness is not relevant to responsible action. (Just like the claim that reading Kant is not necessary for doing good philosophy is a weaker claim than the one that reading Kant is not relevant to doing good philosophy.) Therefore, even though in the previous chapter, I have accepted that being reasons-responsiveness is not always required to performing a responsible action, in this chapter, I bear the burden to show that being reasons-responsiveness is at least *relevant* to performing a responsible action. And this task also helps to answer a question left in the last chapter—what the relationship between responsible agency and responsible action is.

### 4.1 Two models of causal explanation

In the last section, I have provided some hints about why we tend to think that reasons-responsiveness is not explanatory for the specific action. That is, reasons-responsiveness is ostensibly not explanatory because it is taken as an unexercised ability or an unmanifested disposition. This answer is incomplete, however. The conclusion is also based on a specific model of causal explanation, to which I refer as the strict model of causal explanation.

> **The strict model of causal explanation**: To explain an event E is to cite the relevant events/properties which has a causal influence on E.[133]

With this model, we are in a better position to see why an unmanifested modal property does not explain. An unmanifested modal property neither has a causal influence nor corresponds to any events figure in the causal history. For example, the fragility of the glass only has a causal influence in the causal history when it is manifested, if the glass falls to the ground but fortunately, it does not get broken, then intuitively, the property of fragility cannot explain why the glass does not get broken. Likewise, in the case of Frank's shooting Furt, Frank is reasons-responsive in performing the action. That is to say, Frank has an ability to refrain from shooting Furt. Since in the actual history, Frank does not exercise this ability, it does not have a causal influence on the outcome.

However, the strict model of causal explanation is too demanding. Sometimes we explain the occurrence of events without citing any elements with causal influences. Consider the following question: why is David able to solve this complicated mechanics problem so quickly? One can answer this question by saying that he was awarded a degree in engineering at MIT two years ago. Note that the event of being awarded a degree *per se* has no causal influences on the outcome of solving that particular mechanics problem. Nevertheless, it is still explanatory. This suggests that the strict model is inadequate to accommodate all the practice of explanation.

There is a liberal account of causal explanation developed by David Lewis. [134]

> **The Liberal Model of Causal Explanation:** to explain an event E is to provide some information about its causal history.

---

[133] This characterization is neutral to the metaphysics of causation. Different metaphysical theories will count different ontological category as the relate of causal explanation (e.g., facts, events, properties). In addition, different metaphysical theories will have different definition of what causal influence means (e.g., counterfactual theory, manipulation theory, conserved quantity theory).

[134] Lewis 1986.

This model is liberal because it not only allows elements with causal influence to be explanatorily relevant; rather, anything can be explanatory if it provides information about the causal history of the event to be explained. In the example of explaining David's solving the mechanics problem, David's being awarded a degree *per se* has no causal influence on the event of solving that particular mechanics problem. However, it does point to something which probably makes a causal contribution to the explanandum, say, his relevant competence and his training in engineering. Thus, even though the explanans cited is not causally operational, it does provide information about the causal history of the event to be explained.

Lewis's liberal model can account for the explanatory relevance of some particular elements which cannot be handled by the strict model of explanation.[135] I propose that under this model we can also account for the causal explanatory relevance of reasons-responsiveness for this notion does provide important information about the causal history of actions. With this model in hand, I show how we can respond to the Challenge of Irrelevance according to which reasons-responsiveness is not part of the causal history, and not explanatorily relevant.

### 4.2 Reasons-Responsiveness and Causal Information

Now the question is, what information does the notion of reasons-responsiveness provide in accounting for the agent's action? Typically, agents' actions are causally explained by his motivational mental states such as beliefs and desires.[136] These motivational mental states constitute a relevant part of the causal history leading to the actions. In what follows, I will propose that reasons-responsiveness is explanatory because it provides information about the causal history of the action. More specifically, it helps to guarantee that agents' actions are causally explained by the right motivational mental states, namely, the mental states that are sensitive to reasons. The core idea is that reasons-responsiveness is not a psychological faculty over and above the agent's motivational mental states. To illustrate how this idea helps to tackle the challenge of Irrelevance, we should look back to a key move in Sartorio's argument—the case of Frank and insensitive Frank. With this case, Sartorio wants to establish the possibility that two agents share the same actual causal histories while having different capacities of reasons-responsiveness

[135] For example, Beebee (2004) uses this model to account for the explanatory relevance of absence. Jackson and Pettit (1990) use this model to account for the explanatory relevance of high-level dispositional properties which are not causally efficacious in the underlying physical causal processes.
[136] Here I just grant the causal theory of action. I will defend this theory in more details in the following chapters.

in performing their actions. As I will show, once we make explicit the relationship between reasons-responsiveness and the mental states that motivate the action, this case turns out to be incoherent.

Given that Frank and insensitive Frank are stipulated to have similar causal histories with respect to their actions, it is reasonable to imagine that Frank and insensitive Frank's moves result from similar mental states, say, both of their actions result from certain beliefs: Frank holds the belief B #Furt should be killed#. Insensitive Frank holds the belief B* # Furt* should be killed#. B and B* are presumed to be the same type of beliefs. Otherwise, Frank and Insenstive Frank would not have similar causal histories in their actions. However, once we take into consideration that one's ability to respond to reasons is built on one's mental states, it becomes dubious that B and B* are actually the same.

In Sartorio's stipulation, Frank is reasons-responsive in the sense that if Frank learns that Furt is the father of five children who depend on him to survive, Frank will change his mind of killing Furt in light of that reason. However, reasons-responsiveness is not built on nothing. Arguably Frank is responsive to reasons partly in virtue of his mental states being sensitive to reasons. Particularly, his belief B #Furt should be killed# is sensitive to reasons such that it is disposed to be revised in accordance to potential reasons or new evidence: we can imagine that when Franks informs that Furt is critical for his five children to survive, he will hesitate and reflect on his original plan of shooting Furt. This deliberational process is actually subsumed by his mental states revising in accordance with new input—his belief B tends to fade away because of R. This leads him to giving up his plan to kill Furt. Mutatis mutandis, Insensitive Frank is not responsive to reasons mainly because his mental states are not sensitive to reasons. Specifically, even if insensitive Frank finds out that Furt* is the father of five children who depend on him to survive, this fact has no impact on revising his original belief B* for it is insensitive. That is to say, B* will remain and continue to cause Insensitive Frank to implement his plan.

Now the question is, given that B is sensitive to reasons while B* is not, is it plausible to insist that B and B* are the same? I think the answer is not. In particular, we should be cautious about a conflation of belief-talks: when we speak of beliefs, sometimes, we speak of the propositional content of beliefs; sometimes we refer to beliefs as the particular mental states. Since here we are talking about the causal histories which lead to the agents' actions, so I think here beliefs as mental states should be the target of discourse. With this distinction in hand, it is quite plausible that though B and B* are different mental states even though they have the same propositional content.

A popular approach to analyze mental states is the functionalism approach. This approach takes the concept of belief as a functional concept such that a particular belief is defined by its actual and potential

causal connection with other mental states and behaviour.[137] Arguably, being sensitive to reasons is a functional property that helps to define a particular belief because being sensitive to reasons is by nature a set of potential causal connections with other mental states. If a belief is sensitive to reasons, it just means that this belief tends to be revised when it is spoken against by other beliefs. Thus, within the functionalist framework, B and B* are different mental states for they have different functional properties.

A potential objection is that being sensitive to reasons is just a marginal feature of belief so it is not sufficient to mark the difference B and B*. This is not the case. Specifically, if a belief is not sensitive to reasons, then it seems to be a pathological doxastic state. Empirical studies have shown that certain kinds of delusion result from pathological doxastic mental states which are not sensitive to reasons. For example, people who suffer from Capgras delusion will believe that one of their friends or relatives are replaced by an impostor who just looks like their friends or relatives. The patient of Capgras delusion cannot revise their beliefs that one of their friends or relatives is an imposter no matter what evidence has been given to show that is not the case. Likewise, if B* will not be withheld by Insensitive Frank no matter what information is put in, then B* seems to be a pathological doxastic state, rather than a normal belief.[138]

In summary, since insensitive Frank is not responsive to reasons, his belief B* which causally explain his action must be insensitive to reasons and pathological. So B* must be different from B. And if B and B* which respectively explain Frank and insensitive Frank's actions are different mental states, then the underlying causal mechanisms which respectively issue in Frank and insensitive Frank's action must be different. Thus, Sartorio has no ground to assume that Frank's action and insensitive Frank's action have exactly the same causal histories. The case of two agents being different in reasons-responsiveness while having the same causal histories of actions is fundamentally incoherent.

I have suggested that reasons-responsiveness can be shown to be explanatorily relevant to the action within Lewis's liberal model of explanation. This approach requires us to answer the question of what information does reasons-responsiveness provide to the causal history. Here is my answer. The property of being reasons-responsive is a guarantee that the action is issued from the *right* mental states. By saying the agent performs the action in a reasons-responsive way, two pieces of information are provided with respect to the causal history. First, the action is causally explained by the mental states which typically

[137] For more details, see Schwitzgebel (2019).
[138] There is another way to cash out the difference of the two beliefs without relying on functionalism. One might think that the belief B* has a suppressed content 'Furt* deserves to be killed *no matter what*'. B doesn't have this suppressed content (or has different suppressed content) that distinguishes them. I am indebted to Jules Holroyd for this point.

figure in the action explanation, such as beliefs, desires and intentions; and second, these mental states are sensitive to reasons. What I want to propose here is that reasons-responsiveness is not an independent psychological faculty which directly explains an agent's action. Rather, reasons-responsiveness should be (though maybe not exhaustively) understood as a feature of (or realized by) agents' right mental states which causally explain their actions. If the picture I presented so far is convincing, then there is no tension between the actual-sequence view with the modal view of reasons-responsiveness. Nor is there any tension between the Irrelevance Principle (IP) and the modal view of reasons-responsiveness. That is to say: the actual sequence involves a causal history that makes reference to the reasons-sensitive (or reasons-insensitive) mental states of the agent. These have an explanatory role in producing the action, and are not irrelevant.

This account fits well with our intuitions regarding moral responsibility. For example, a drug addict is not reasons-responsive in taking the drug and because of that not responsible for the behaviour. Here being not reasons-responsive provides the information that the agent's action is not motivated by normal beliefs and desire, but probably an irresistible urge.

**Concluding Remarks: The Revenge of Compatibilism**

The Frankfurt-Style cases usually are taken as an important challenge to the incompatibilism and because of that a support to the compatibilism. However, as I have shown in the last chapter and this chapter, the Frankfurt-Style cases equally pose challenges to compatibilists who rely on a modal analysis of free agency, particularly, the reasons-responsiveness theory. Specifically, the Frankfurt-Style cases challenge the reasons-responsiveness theory in two different ways. First, according to the intuitive judgment of the Frankfurt-Style cases, the ability to do otherwise—even understood conditionally—is not necessary for moral responsibility. And second, according to the rationale motivated by the Frankfurt-Style cases, the conditional ability to do otherwise or the agent's reasons-responsiveness is irrelevant to the explanation for the action and *ipso facto* irrelevant to the moral responsibility. I diffuse the first worry in the last chapter (by distinguishing responsible agency and responsible action and arguing that the former is not necessary for the latter) and the second worry in this chapter (by showing the explanatory relevance of reasons-responsiveness to responsible action). So far, I have shown that both the intuitive judgment and the plausible rationale (the Irrelevance Principle) from the Frankfurt-Style cases are compatible with the reasons-responsiveness theory.

Since I have shown that the Frankfurt-Style cases do not cause trouble to compatibilism, it is worth asking whether the Frankfurt-Style cases lend some support to compatibilism. I think the answer is yes.

The Frankfurt-Style cases favour compatibilism, but less straightforwardly. According to the Irrelevance Principle motivated by the Frankfurt-Style cases, factors have to be explanatorily relevant in order to be relevant to responsible action. The challenge to the reasons-responsiveness theorists is to provide a story about how reasons-responsiveness, as an unmanifested modal property, is explanatorily relevant. I have provided such a story with Lewis's model of causal explanation. That is, the notion of reasons-responsiveness helps to pick up the right mental states which constitute the causal history of the action. A similar challenge confronts incompatibilists who hold that the categorical ability to do otherwise is necessary for moral responsibility. Given that the categorical ability to do otherwise is also an ability unexercised, it is not clear how it is explanatorily relevant. More importantly, incompatibilists cannot use the same strategy as compatibilists do. This is because, the categorical ability to do otherwise does not provide any information about the causal history leading to the action.[139] Thus, on the foothold of the Frankfurt-Style cases, the compatibilists do gain a dialectical advantage over the incompatibilists: compatibilist freedom can be shown to be explanatorily relevant while incompatibilist freedom is not.

I have now developed a complete account of reasons-responsiveness. This account includes two basic ideas. First, to possess responsible agency is to be reasons-responsive, which consists of the conditional ability to do otherwise (this idea is defended in the last chapter); Second, reasons-responsiveness can figure in the causal explanation for the responsible action by providing information about the causal history of action. This second idea answers the question left in the last chapter—what the relationship between responsible agency and responsible action is. It also paves the way for integrating the reasons-responsiveness theory with a theory of action. That is, to exercise reasons-responsiveness is to perform an intentional action in a reasons-responsive way, which can be further cashed out in terms of causal interactions involving the motivating mental states which are sensitive to reasons.

Some may have noticed that my account of reasons-responsiveness hinges on a specific theory of action, namely, the causal theory of action. Very roughly, this theory claims that action can be reduced to causal interactions between motivational mental states and bodily movements. This theory, of course, calls for defence; a topic I will turn to from the next chapter.

[139] Of course, by pointing out that the agent has the ability to do otherwise in a categorical sense, we do know something about the causal history—that is, the causal history is indeterministic. However, this information is not about any factors which make causal contributions to the happening of the action. Put it differently, simply knowing that the causal history is indeterministic does not improve our understanding for the question why that particular action occurred. Thus, it is still difficult to see the explanatory relevance of indeterminism.

## Chapter 4: The Problem of Causal Deviance

**Abstract:** In this chapter, I tackle the problem of causal deviance. In doing so, I develop a novel account of control. This account takes into considerations several factors which may influence our judgments of whether a certain object is in control. These factors include causation, purposiveness, accuracy, reliability and flexibility. According to this account, our conception of control is multi-faceted and control is a notion that comes in degrees. An implication is that it is neither possible nor necessary to draw a sharp line between the deviant causal chains and the non-deviant causal chains. In effect, the problem of the causal deviance is dissolved.

### 0. Introduction

In the previous two chapters, I have developed an account of reasons-responsiveness. The account has two basic ideas. First, reasons-responsiveness consists of the conditional ability to do otherwise. Second, reasons-responsiveness, as a modal property, is explanatorily relevant to the agent's free actions. Specifically, reasons-responsiveness can provide information about the mental antecedents which causally explain the action. This idea presumes a specific type of action theories, namely, the causal theory of action. To have a more complete picture of agency, the causal theory of action should be defended and incorporated to the reasons-responsiveness theory.

I take the causal theory of action (or CTA henceforth) as a group of theories which aim to explicate the nature of action in terms of causal interactions between specific mental states and the bodily movements.[140] Specifically, these theories provide elegant answers to the question of what distinguishes actions from bodily movements which merely happen to us, e.g., trembling, knee-jerk reflexes. For example, according to a popular version of CTA, known as the 'standard story of action'[141], actions are bodily movements caused by agent's motivating mental states such as desires, beliefs, intentions, etc.[142] I call this the standard causal theory of action (or standard CTA henceforth).

The standard CTA became popular because of the naturalistic climate in philosophy. According to this account, action is embedded into an event-causal framework rather than an agent-causal one. This is an

[140] Understood in this way, the Causal Theory of Action is more ambitious than just providing an account for the action explanation. More on this point in section 3.
[141] This term is coined by Velleman (1992).
[142] Note that CTA here does not refer to a single theory but a set of theories. The details of the mental antecedents that cause the bodily movement depends on the specific theory one commits to.

advantage since many philosophers believe that the picture revealed by sciences is an event-causal one. That is, the relata of causal relations must be events. If actions are parts of the natural phenomena that could be studied by sciences, then actions must occur within an event-causation framework. In this sense, standard CTA is a promising project to naturalize action and agency for it uses no more than event-causal notions to characterize action. By comparison, an agent-causation theory of action is not friendly to naturalism in the sense that it involves the notion of agent-causation which cannot fit into an event-causation framework and in effect cannot fit into the picture provided by sciences.

However, the standard CTA notoriously faces the problem of causal deviance, which makes many philosophers doubt whether actions can be captured by an event-causal framework. Specifically, there are counterexamples in which the agents' bodily motions are caused by their motivating mental events but intuitively we will not take their bodily motions as actions. Consider the famous case introduced by Donald Davidson:

> **Climber**: A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never chose to loosen his hold, nor did he do it intentionally. (Davidson 1973/2001, 79)

In this case, though the climber's loosening his hand is caused by his mental events or states (belief and want), intuitively, we would not regard this move as the action done by the climber.

Regarding the deviance cases, defenders of CTA add that to constitute an action, the agent's bodily movement should be caused by her mental states in a non-deviant way. Several approaches have been proposed to analyse the notion of non-deviant causal chain, but none of them manages to convince the opponents of CTA—those analyses, though succeed to handle the classic cases of causal deviance, are all subject to more sophisticated cases of deviance. In addition, some of the revised versions of CTA may involve too strong conditions that may exclude genuine cases of action. In a word, it seems to be extremely difficult for the defenders to provide a set of sufficient and necessary conditions for non-deviant causal chains.

If CTA cannot successfully solve the problem of causal deviance, then the next question is, so what? Opponents of CTA are glad to announce that the project of reducing action into event-causal interactions has completely failed. However, this conclusion is too hastily drawn. In this chapter, I will argue that CTA does not have to provide a set of sufficient and necessary conditions for the non-deviant causal chains. I will defend a deflationist position of action, according to which we cannot and should not

require a set of sufficient and necessary conditions for action. Specifically, I will argue that the fact that the sufficient and necessary conditions for action cannot be provided is not because that action cannot be embedded into an event-causation framework; rather it is because our ordinary judgements about control come in degrees.

Here is the agenda for this chapter. In the first section, I will set the background for the debate and provide an approximation of the desiderata for a successful solution to the problem of causal deviance, which includes the requirement of sufficiency and necessity. In the second section, I will use the sophisticated CTA as an example to show how the requirement of sufficiency and necessity occupy a central role in the debate. In the third section, I entertain Davidson's approach to rejecting the requirement of sufficiency and necessity. In the fourth section, I will provide my own approach to rejecting the requirement of sufficiency and necessity. Specifically, I will develop an account of control, to which I refer as the gradualist conception of control. This account shows that our ordinary conception of control comes in degrees. Since there is no sharp line between control and non-control, there is no sharp line between deviance and non-deviance of causal chains. I conclude that an adequate theorization of non-deviant causal chains is neither possible nor required by the defenders of causal theories of action.

## 1. Some Clarifications for Causal Deviance

### 1.1. Basic Deviance and Non-Basic Deviance

As mentioned in the introduction, philosophers tend to reach a consensus that the conditions specified by the causal theories of action or CTA are not sufficient to define action because of cases of causal deviance. In addition to the Climber composed by Davidson, there are some other deviance cases which are often cited in the literature.

> **Shooter:** A man may try to kill someone by shooting at him. Suppose he misses him, but the shot stampedes a heard of wild pigs which tramples the intended victim to death. (the case is devised by Dan Bennett, quoted from Davidson 1973/2001, 78)

> **Robber:** Imagine a robber who wants to tell his confederates that the time has come to start a robbery. He is at a party and believes that by spilling the contents of his glass he will signal that the time to start the robbery has arrived. He then forms the intention to spill the contents of his glass. However, because he is an inexperienced robber, his forming this intention causes his hand to tremble and the contents of his glass are spilt. Although the outcome of

> this causal chain of events is exactly what the robber intended, it is not one of his intentional actions. (Frankfurt 1978, 157, paraphrased)

In these cases, the behaviour of the agents are all caused by their mental states which are supposed to rationalize their actions. However, we do not think that in these cases the agents act intentionally.

Many philosophers agree that the causal deviance chains can be categorized into two different types: the non-basic deviance and basic deviance. This distinction can be made in several ways. Before making the distinction, a notion of basic action should be introduced. In philosophy of action, the notion of basic action refers to an action which the agent performs via no other means. To put it differently, a basic action cannot be performed by doing anything else. For example, moving my arm is a basic action, while hailing a cab is a non-basic one. With this notion in hand, we can see how the basic causal deviance differs from the basic one.

First, in the cases of basic deviance, the deviance occurs in the causal connection between the agents' mental states and the "basic actions" while in the cases of non-basic deviance, the deviance occurs in the causal connection between the" basic actions" and the corresponding effects.[143]

Second, in the case of non-basic deviant causal chains, the behaviour of agents are not directly caused by the mental states which are supposed to rationalize their action; by comparison, in the cases of basic deviance, the behaviour or bodily movement are directly caused by the mental states which are supposed to rationalize the agents' actions.

From these points, the case of Climber and Robber can be categorized as basic deviance; the case of Shooter can be categorized as non-basic deviance.

Many philosophers agree that there is a direct solution to the case of non-basic deviance: those cases go deviant because the events do not unfold following the action-plan which is represented by the mental states of the agent. To exclude non-basic deviance, we just need to set up the conditions that the agent's movement are caused by the mental states which represent the action-plan.[144] This strategy, however, cannot apply to the cases of basic deviance because the deviances in such cases occur in between the mental states representing the action-plans and the bodily movements. In addition, in the cases of non-basic deviant causal chains, there are actually intentional actions being performed (at least there are intentional actions under certain descriptions); by comparison, in the cases of basic deviant causal chains,

[143] Since in the cases of deviance actions are degenerated, I use quote marks to indicate that the actions here are not actions in its full-blown sense but just bodily movements or behaviour which are supposed to be intentional action in normal cases.
[144] See e.g., Bishop (1989).

no intentional actions are being performed. That is to say, the non-basic deviance only undermines action under certain descriptions while the basic deviance undermines action *per se*. I suggest that the cases of basic deviance constitute a real challenge to the causal theory of action for these cases purport to show that action by nature cannot be reduced to purely event-causation terms. Since the cases of basic deviance are the really difficult ones and pose challenges to the nature of action, the following discussion is restricted to the cases of basic deviance.

### 1.2 Is the Problem Empirical or Philosophical?

Some defenders of CTA suggest that the problem of causal deviance is to be solved by empirical sciences (especially neurosciences) rather than philosophy (Goldman 1970; Moore 2010). Here is a quote from Goldman, who is the often-cited defender of this approach.

> Precisely what is this 'characteristic' mode of causation by which wants and beliefs cause intentional action?... A complete explanation of how wants and beliefs lead to intentional action would require extensive neurophysiological information, and I do not think it is fair to demand of a *philosophical* analysis that it provide this information…a detailed delineation of the causal process that is characteristic of intentional action is a problem mainly for the special sciences. (Goldman 1970, 62, original emphasis)

The motivation for this approach is well-founded. Philosophers probably do not have either the information or vocabularies to specify the non-deviant causal chains. The resources which philosophers usually invoke are the concepts of folk psychology, such as beliefs, desires, intentions, wants, volitions and the like. However, folk psychology is not a rigid science; it does not provide an accurate and complete description of the mental processes underlying the occurrence of action. Thus, it seems to be reasonable to expect that the analysis of non-deviant causal chains should be provided by scientists. Waiting for scientists to solve the problem, however, does not genuinely solve the problem. Before the development of molecule theory and the empirical investigation, people did not know that the concept of water referred to a unified natural kind, namely, $H_2O$. Similarly, it is an open question of whether there are uniformly causal processes corresponding to the ordinary concept of action. In addition, there are certain philosophical worries that no uniform causal process between mind and body corresponding to intentional actions can be established (e.g., Davidson's anomalous monism).[145]

[145] More on this point in section 3.

But even presuming that scientists will eventually discover uniform causal processes correlated with intentional action, this cannot be a solution to the problem of causal deviance. As said in the introduction chapter, an ideal version of CTA needs to meet two requirements—the explanatory requirement and the ontological requirement. Recall the two requirements are as follows. To meet the explanatory requirement, the CTA need to facilitate our understanding of how action can arise through the causal processes provided by the analysis. To meet the ontological requirement, the CTA need to show that action is nothing more than the causal process characterized by the analysis.

Can an empirical solution meet the explanatory requirement? Let's suppose that there is the same kind of neural process underlying all intentional actions and that this neural process sufficient to exclude all behaviour resulting from deviant causal chains. Still, it is probable that we cannot understand why action arises from such neural processes. The point is not that sciences cannot contribute to our understanding of non-deviant causal chains at all. Rather, the point is that there is a certain normative dimension of the problem of causal deviance, which cannot be settled merely by empirical methods.[146] As I will show in the following section, the key to solving the problem of causal deviance is to show that how agentive control can arise from an event-causal process, which has to take into account certain normative considerations.

Likewise, an empirical solution cannot meet the ontological requirement either. As Bishop (1989) convincingly points out, a process correlating with the occurrence of action is not necessarily a process constituting the occurrence of action.[147] While science at best can point to certain correlations between action and certain causal processes, according to Bishop, the aim of CTA is to show that action can be genuinely constituted by event-causal processes.

### 1.3 Desiderata for A successful Solution: A First Approximation

If a solution to the problem of causal deviance cannot be purely empirical, then what should be the desiderata for a successful solution to the problem? A successful analysis of a non-deviant causal chain must be part of the whole analysis of action. Thus, whether the problem of causal deviance can be solved depends on whether a successful CTA analysis can be provided. We can first look at an influential

[146] As Keil (2007, 75) puts it, "[m]other natures draw no distinction between normal and deviant causal chains". To illustrate this point, Keil helpfully introduces an analogy. That is, the distinction between effects and side-effects of a medicine. There is distinction of effects and side-effects of a medicine because some of the effects are desired by the human for the usefulness of curing diseases while others not. This distinction is mainly marked by human's interests rather than nature.

[147] Bishop (1989, 96–97). See also Mayr (2011, chap. 5).

proposal suggested by Bishop. As mentioned, Bishop holds that a CTA analysis aims to show that action is constituted by rather than correlated with an event-causal process. According to Bishop, the relation of constitution can be specified by a modal stipulation: if action is constituted by a certain event-causal process, then it will be *inconceivable* for an action to occur without such an event-causal process in any possible worlds with similar ontology and causal order as the actual world. Bishop thinks that adding this modal stipulation to the CTA analysis can guarantee that action is not just contingent on the event-causal process depicted in the analysis, but genuinely constituted by it. As I will show in the following section, this modal stipulation does not serve this aim. But currently, I will put this dispute aside. I will temporarily follow Bishop's proposal and list several desiderata for a successful CTA analysis.

First, the CTA analysis should only invoke notions recognized in an event-causal framework. Since the key idea of CTA is that the phenomena of action (which typically thought to be agent-causal) are not over and above the event-causal process, a successful CTA analysis should not make reference to notions with an agential residue, such as agent-causation, control, guidance, and the like. For such notions are all conceptual neighbours of action or agency, which is the target of analysis, thereby making the analysis circular. Call it the requirement of *non-circularity.*

Second, the CTA analysis must apply to all possible worlds with similar ontology and causal order to the actual world. The aim of this modal stipulation, as said, is to guarantee that action is not just correlated with but constituted by the causal process depicted in the analysis. Call it the requirement of *modal robustness*.

More importantly, for Bishop and many others, the analysis must provide a set of sufficient and necessary conditions for action. This means that the analysis will be able to rule out all possible cases of causal deviance but not to exclude cases which we intuitively think to be action. Call it the requirement of *sufficiency and necessity*.

Actually, providing sufficient and necessary conditions for action becomes the holy grail for many defenders of CTA. Most debates centre upon whether the analysis in question is sufficient and necessary for action. As to be shown in the following section, defenders of CTA fail to convince their opponents mainly because the requirement of sufficiency and necessity is not met. However, in the literature, there has been little reflection on this requirement. Why must a successful CTA meet this requirement? Of course, opponents of CTA would probably reply that the failure to meet this requirement implies that action cannot be reduced into an event-causal process. That is, the failure can only be explained by the impossibility of reducing action into an event-causal process. One of the main tenets of this chapter is to rebut this requirement. I will argue that the task of CTA is not to seek the sufficient and necessary

conditions for action. Rather, it is to provide a reductive explanation for action. This reductive explanation involves a functional analysis of agentive control. Based on this functional analysis of agentive control, I will show that a sufficient and necessary conditions for action should not be expected.

Once liberated from the requirement of sufficiency and necessity, we can show that CTA does have resources to solve the problem of causal deviance. The resources I rely on are found within the sophisticated CTA (mainly developed by Peacocke (1979) and Bishop (1989)) which involves the sensitivity condition and feedback loops condition. Now I turn to these two conditions.

## 2. The Sophisticated Version of CTA

According to the standard CTA, actions are bodily movements caused by the agent's motivating mental states such as desires, beliefs, intentions, etc. Because of the causal deviance cases, defenders of CTA come to realize that merely invoking the causal interactions between the body and the motivating mental states is not sufficient to distinguish actions from other bodily movements. Defenders add that when the action occurs, there should be specific structure underpinning the event-causal process. This is how the conditions of sensitivity and feedback loops come into play.

### 2.1. The Sensitivity Condition

From the case of Climber and the case of Robber, we can see that a causal chain becomes deviant because an agitation state arises (e.g., nervousness) between the original mental states (e.g., desires, intentions) and the resultant bodily movement. If we can rule out this agitation state, then we can probably have a non-deviant causal chain. One approach to ruling out the deviance cases is to close the gap between the original mental states and the bodily movement by adding intermediate mental states (such as a proximal intention) to the causal chain.[148] There is an obvious problem with this approach: in principle, an agitation state can still arise between the intermediate mental states and the bodily movement. Thus, instead of introducing more new intermediate mental states, a better approach is to focus on the characteristics of the causal relation between the original mental states and the resultant bodily movement. Specifically, many defenders of CTA held that when an action takes place, the resultant bodily movements must be sensitive to the mental states.[149] A straightforward way to specify the

[148] E.g., Brand 1989; Mele 1992; Searle 1983.
[149] E.g., Peacocke 1979; Bishop 1989; Smith 2009.

idea of sensitivity is by counterfactual characterization: to say that the resultant bodily movement is sensitively caused by the mental states is just to say that if the content of the mental states changes slightly, the bodily movement will change accordingly. Suppose an agent intentionally raised his hand to reach the bottle in a specific way at a specific moment. Had the content of his intention changed slightly, say, to raise the hand at a different moment, or in a different direction, his bodily movement would correspondingly occur at a different time, or in a different way.[150]

We can apply this approach to analyse how the agent's behaviour goes deviant in the case of Climber. In this case, the agent's mental states first cause a state of nervousness which happens to cause his loosening of the hand. Since the nervousness intrudes the causal chain, is reasonable to expect that his motion is not sensitive to the contents of his mental states simply because the state of nervousness is blind to the contents of his mental states. Suppose in a counterfactual scenario the agent also acts out of a state of nervousness but at this time he changes his contents of desire a little bit, say unfolding his hands at a different speed or in a different gesture, his motion would not respond to this change and his hand would move in the same way as in the actual case.

However, there are worries that the sensitivity proposal cannot rule out deviant causal chains thoroughly. Peacocke himself has devised a convoluted counterexample, which involves two agents: It is the first agent who forms the relevant mental states and intends to act. However, the causal chain is blocked and preempted by the second agent, a knowledgeable and resourceful neuroscientist who can read the first agent's mind precisely and manipulate the first agent's brain activity. Thus, whenever the first agent forms an intention to act, the neuroscientist will block the natural causal link from the intention to the bodily movement; in the meantime, the neuroscientist will read the intention from the first agent and stimulate the agent's motor system, causing the first agent to move accordingly.[151] Note that the first agent's body motion is perfectly sensitive to the contents of his mental states because of the second agent, who will know exactly the contents of the first agent's mental states and manipulate the first agent to behave exactly in accordance with the contents. However, we seem to have an intuition that the first agent does not perform any actions. This suggests that merely by invoking the sensitivity condition is not sufficient to rule out all deviant cases.

[150] Apart from the counterfactual interpretation, Peacocke (1979) propose another way to interpret the idea of sensitivity, which is known as differential explanation. According to Peacocke, to say that the bodily movement is sensitive to the mental states is to say that the bodily movement can be differentially explained by the mental states. This relation of differential explanation grounds on a certain law of nature, which can be specified by a certain function.

[151] See Peacocke 1979, 87.

Bishop (1989) calls the cases involving two agents the heteromesial cases. He is not satisfied with Peacoke's solution to the problem by adding qualification to the sensitivity condition: that is, the causal chain from the agent's intention to the resultant bodily movement should not go to another agent. For this solution suggests that any heteromesial cases are deviant. Bishop holds that heteromesy 'sometimes introduces deviance, sometimes not—depending on the precise nature of the second agent's participation in the causal chain'.[152] To illustrate this point, he introduces a non-deviant heteromesial case. An agent's neural system is largely replaced with a prothesitc system, a device external to the brain. One day, when the agent tries to act, one of the wires in the system gets broken. Fortunately, another agent happens to pass by, assist the first agent to complete his action by resoldering the wire. Bishop holds that contra Peacocke's preemtive heteromesial case, this assistance heteromesial case is intuitively non-deviant despite the causal chain involving another agent's intention. But what makes the difference between these two cases? This question motivates Bishop to introduce a second condition for non-deviant causal chain—the feedback loops condition.

### 2.2 The Feedback Loops Condition

In order to accommodate the problem triggered by the heteromesial cases, Bishop proposes that in a non-deviant causal case, the mental states not only cause the bodily motion in a sensitive way, but also cause it sustainedly. The idea of incorporating sustained causation into the CTA analysis is first introduced by Thalberg (1984), who suggests that when an action occurs, the causal interaction between the mental states and the matching behaviour constitutes an ongoing causal process. Specifically, the mental states not only initiate the matching behaviour, but also guide and regulate it. However, Bishop is not satisfied with this proposal because Thalberg's analysis still involves agentive notions such as guiding and regulating, which are just conceptual neighbours of action. In other words, Thalberg's analysis does not meet the requirement of non-circularity.

To show that sustained causation can be realized by an event-causal process, Bishop borrows the lesson from the servosystems studied in cybernetics. A servosystem obtains control if it produces or maintains outputs which are within the intended range of values. Specifically, it obtains control in virtue of a mechanism known as feedback loops. The mechanism of feedback loops enables the system to continuously monitor the outputs and adjust them to the intended values; if it turns out that the actual performance goes deviant, the system will make a revision to guarantee the intended output. For example, a thermostat is a simple servosystem with feedback loops. A thermostat controls the indoors temperature

[152] Bishop 1989, 158.

by continuously monitoring the actual temperature and comparing it to the pre-set value. If the actual temperature is higher, then the cooling mechanism will lower down the temperature; or vice versa, the heating mechanism will work.

Bishop holds that the agent exercises control through her action basically in a similar way to a servosystem. He suggests that the agent's mental process involved in her action can be viewed as a control system which is structurally similar to a servosystem. Specifically, the control system of the agent continuously receives feedback of the matching behaviour and regulate it accordingly. Though Bishop says very little about what the control system of the agent looks like, the take-home message is that since the servosystem can be described by purely event-causal processes, so can the agent's control system.

Besides, we can outline several features of this feedback loops approach by comparing it with the standard CTA approach. Recall that according to the standard CTA approach, when an action takes place, the bodily movement is caused by specific motivating mental states. The causal process depicted by the standard approach is basically linear and of a single level. By comparison, according to the feedback loops approach, the causal process is more dynamic and of multiple-levels. Specifically, since this control system plays the role of monitoring and adjusting the matching behaviour, it must consist of a set of mental states including not only motivating mental states but also perceptual and proprioceptual states.[153] In addition, these mental states operate on both personal and sub-personal level. The personal level mental states, such as intention, only represents the action in its rough outline; while the fine-grained details required by the control system are represented by sub-personal mental states.[154]

In summary, Bishop believes that with the sensitivity condition and the feedback loops condition we can provide a sufficient and necessary analysis for action. That is, when the action occurs, i) the agent's mental states should cause the bodily movement in a sensitive way; ii) if the causal process involves feedback loops, the feedback loops' signals about the behaviour should go back to the agent's own control system.[155] Call the CTA involving the condition of sensitivity and feedback loops the sophisticated CTA. The analysis of sophisticated CTA can explain why the preemtive heteromesial case is deviant in a simple way: in that case, the feedback signal goes to the neuroscientist's control system instead of the agent's in question.

There are some qualifications with Bishop's sophisticated CTA analysis. First, the feedback loops condition is specified in a conditional, which means that feedback loops are not necessary for actions.

[153] For the importance of perception and proprioception in action, see Wu (2016).
[154] For the hierarchy representations of action, see Pacherie (2012).
[155] See Bishop 1989, 172.

Bishop thinks that there are some actions which take too little time to have feedback signals. In those cases, Bishop holds that the sensitivity condition alone is sufficient for action. Second, the feedback loops condition requires that the feedback signal return to the agent's own control system. Some philosophers worry that the reference to agent violates the requirement of non-circularity for we cannot have a theory of ownership of agent unless we have a proper understanding of agency.[156] However, it is reasonable to expect that an account of individualization of agents can be developed regardless of an account of action. In particular, we can develop an account of individualization of agent with recourse to the personal identity literature, which is largely independent of the discussions about action and agency.

### 2.3 Counterexamples to the Sophisticated Analysis

The sophisticated version of CTA is the most defensible CTA in the current literature. Still, opponents find the analysis unsatisfactory for it fails to meet the requirement of sufficiency and necessity. And because of that, they do not think this theory manages to solve the problem of causal deviance.

#### 2.3.1 The sophisticated analysis is not sufficient

Bishop introduces the feedback loops conditions to rule out the deviant heteromesial cases. However, some philosophers have pointed out that even with the feedback loops conditions, the CTA analysis is not sufficient to rule out all deviant heteromesial cases.[157] According to the analysis of the sophisticated CTA, the preemptive heteromesial cases are deviant because in this case, the feedback loops signals are returning to the neuroscientist instead of the agent. However, we can easily tailor the preemtive heteromesial case to meet the condition of the analysis: we can imagine a preemptive heteromesial case in which the feedback information not only routes back to the neuroscientist but also to the central process of the agent. That is, the feedback information is shared with the neuroscientist and the agent. Since there is no significant difference between this tailored case and the original one, it is unclear how the introduction of the feedback loops solves the problem. One may want to block this objection by stipulating that the feedback loops signals *exclusively* go back to the agent's control system. However, this seems to be an *ad hoc* move if we already admit that heteromesial cases (cases involving two agents) can be non-deviant (as Bishop does).

[156] For example, Steward (2012, 59–60).
[157] For example, Mayr (2011), and Aguilar (2012).

### 2.3.2 The sophisticated analysis is not necessary

Apart from the feedback loops conditions, the analysis of the sophisticated CTA also set a necessary condition for non-deviant causal chain—the sensitivity condition. However, opponents argue that the sensitivity condition is not necessary for all actions. For example, Mayr (2011, 120) has pointed out that there are certain simple bodily abilities which can be exercised in a single way. For example, one can roll his tongue or wiggle his ear. Nevertheless, he can roll his tongue or wiggle his ear in a simple way. There cannot be any variations in the content of his intention to roll his tongue or wiggle his ear. Mayr refers to these abilities as *on-off abilities* and holds that the on-off abilities do not require sensitivity condition.

Sehon (1997) presents a classical example to show that the sensitivity condition is not necessary. Suppose a baseball pitcher is practising her delivery. She first forms an intention with the content, say, throwing the pitch at a velocity of 70 mph, and then she throws the pitch exactly at that velocity. Now suppose if the pitcher has formed an intention with slightly different content, say throwing the pitch at 69 mph or 71 mph, due to the lack of dexterity, she would still throw the ball at 70 mph. That is to say, the pitcher's movement is not strictly sensitive to the content of her intention. However, we still take the pitcher's throwing as an action.

Lilian O'Brien (2012) also devises a scenario to show that the sensitivity condition is not necessary. She uses a case in which the agent reliably predicts that her particular intention to perform a task will invariably generate a state of agitation and she exploits this state to complete her action. Suppose Maria has a job to let go of balloons from her hand at the ceremony of the Olympics. At the rehearsals, she found that whenever she forms the intention to let go the balloons go, the intention will generate a state of nervousness and the nervousness coincidently cause her hand to let the balloons go. After repeated practice, she learns to live with this nervousness and she uses the nervousness to help herself to complete the task. In this case, because of the nervousness, Maria's motion is not sensitive to her intention. However, O'Brien holds that Maria's motion still constitutes an action.

All of the above cases motivate doubts that the sensitivity condition is genuinely necessary for an analysis of action. Thus, opponents of CTA find the sophisticated version fails to satisfy the requirement of the sufficiency and necessity.

## 2.4 State of The Art

In section 2, I have focused on the sophisticated CTA which involves a sensitivity condition and a feedback loops condition. The sophisticated CTA is defensible. It can handle the classical causal deviance

cases, such as the Climber devised by Davidson. However, for many philosophers, the sophisticated CTA is still an inadequate account of action for it fails to provide a sufficient and necessary analysis for action. In the current literature, several CTA proposals uniformly fall to a similar situation—they can help to circumvent the classical deviant cases, but fail to accommodate the new convoluted ones devised by opponents. Thus, they fail to meet the requirement of sufficiency and necessity.[158]

So how to explain this failure? Opponents hold that this failure indicates that the causal theory of action cannot fully capture the notion of action. I call this the *explanation of anti-reduction*. For example, Wilson writes: "the evidence points to a more than infelicity or incompleteness in the various causalist projects—it points, that is, to a global breakdown in the whole point of reduction" (Wilson 1989, 258). And Yair Levy writes that "the continued failure to vindicate such an analysis merits exploring alternative, arguably more promising, research program" (Levy 2013, 710). In sum, many philosophers think that the causal theory of action fails to handle causal deviance cases mainly because the requirement of sufficiency and necessity cannot be met. And the lack of a sufficient and necessary analysis indicate that the framework employed by CTA is ontologically inadequate to capture the notion of action.

However, is this requirement crucial for a satisfactory CTA? I think not. In what follows, contra the explanation of anti-reduction, I will argue that the lack of sufficient and necessary event-causal analysis for action should not be explained by the ontological defectiveness of CTA. I will present two alternative explanations for the lack of sufficient and necessary analysis for action. The first one is from Donald Davidson, to which I now turn.

### 3. Davidson's Deflationism of CTA

Even though many philosophers believe that a satisfactory CTA needs to provide sufficient and necessary analysis for action, there are some defenders of CTA who are not bothered by this requirement. Specifically, they concede that the CTA does not have resources to provide a sufficient and necessary analysis of the non-deviant causal chains. Nevertheless, they deny that this is a sufficient reason to abandon CTA. Call this position the deflationism of CTA. Donald Davidson holds a deflationist attitude toward a solution to the problem of causal deviance.[159] In different places, he expresses the pessimistic prospect of providing an adequate analysis of non-deviance. For example, he writes:

[158] For a critical review of the major proposals to tackle the problem, see Mayr 2011
[159] Others who hold a similar attitude include Keil (2007) and Tännsjö (2009).

> Can we somehow give conditions that are not only necessary, but also sufficient, for an action to be intentional, using only such concepts as those of belief, desire and cause? I think not. (Davidson 1974/2001, 232)

> Several clever philosophers have tried to show how to eliminate the deviant causal chains, but I remain convinced that the concepts of event, cause and intention are inadequate to account for intentional action. (Davidson 2004, 106)

But why Davidson holds such a pessimistic attitude? This is probably due to his special conception of human mind—that is, even though ontologically the mental is token-token identical to the physical, no exceptionless laws can be generalized in the realm of psychology. This position is now known as anomalous monism. If the psychology is anomalous, then we should expect no sufficient and necessary analysis for a non-deviant causal chain of action. Otherwise, this sufficient and necessary analysis will become a strict psycho-physiological law. Davidson provided the following argument:

> For a desire and a belief to explain an action in the right way, they must cause it in the right way, perhaps through a chain or process of reasoning that meets standards of rationality. I do not see how the right sort of causal process can be distinguished without, among other things, giving an account of how a decision is reached in the light of conflicting evidence and conflicting desires. I doubt whether it is possible to provide such an account at all, but certainly it cannot be done without using notions like evidence, or good reasons for believing, and these notions out run those with which we began. (Davidson 1974/2001, 232)

If I understand this argument correctly, what Davidson means is that for a causal chain to be non-deviant, the mental states causally responsible for the action must meet the requirement of rationality, which are characterized with notions like evidence or good reasons for believing. However, there are no law-like generations over these normative notions. This idea, of course, is closely connected to his argument for the anomalousness of psychology. Unlike the physical realm which is constitutive of causation, the psychological realm is constitutive of rationality. Therefore, no causal laws can be generalized in the psychological realm. Davidson's anomalous monism may be the primary motivation for him to adopt a deflationist attitude toward action and he has independent arguments for his anomalous monism.[160] And his deflationist attitude also supports his anomalous monism:

[160] See Davidson 1970/2001; 1974/2001.

> What prevents us from giving necessary and sufficient conditions for acting on a reason also prevents us from giving serious laws connecting reasons and actions. (Davidson 1974/2001, 233)

To summarize, Davidson's deflationist attitude toward action is based on his anomalous monism and his deflationist attitude motivates his anomalous monism. Actually, because of his view of psychology, he not only gives up on providing a sufficient and necessary analysis for non-deviant causal chains, but also gives up on providing any specification for this notion.

If Davidson's argument works, then perhaps defenders of CTA have a way to reject the requirement of sufficiency and necessity. According to this strategy, there is no successful way to provide a sufficient and necessary analysis for action; while this failure is not due to CTA *per se*. Specifically, Davidson will probably disagree that the intractability of the problem of causal deviance is due to any inadequacy of the event-causal framework. Ontologically speaking, an action still arises from an event-causal process.[161] Action resists a sufficient and necessary analysis just because there are no strict laws that bridge psychology and physiology.

However, I do not think that Davison's deflationism really helps the current CTA defender to reject the requirement of sufficiency and necessity. The first problem with Davidson's deflationism is that it completely hinges on anomalous monism, which is a controversial position. To defend Davison's deflationism, one needs to defend anomalous monism. This seems to be a heavy burden of proof. Besides, even if it is granted that we have sufficient reasons to accept anomalous monism, there is still the question about why Davidson's deflationism is adequate.

That is, without a principled way to distinguish deviant causal chains from non-deviant causal chains, a CTA fails to be a satisfactory theory about the nature of action. When Davidson first proposed a CTA in his classic paper "Actions, Reasons and Causes" (1963), his aim was not as ambitious as his followers. What he tried to do was to provide a theory of action explanation. Specifically, he wanted to show that rational explanation of action is a kind of causal explanation. And his argument can run without any specification of non-deviant causal chains. However, the current defenders of CTA are answering a different question—the nature of action.[162] That is, how actions differ from mere happenings. It is unlikely that one can have an account of action without a proper analysis of non-deviant causal chains.

[161] I think this is what Davidson means when he writes that: "[w]e would not, it is true, have shown how to define the concept of acting with an intention; the reduction is not definitional but ontological." (Davidson 1978/2001, 88)
[162] The explanation of action mainly concerns about why an action occurs; while the nature of action, as I will show in the next section, concerns about how an agent maintains control during an action. These two questions can be

## 4. Control and Causal Deviance

In the last section I have discussed Davidson's deflationism and the problems with it. Still, I am sympathetic to this position. There are two points that I share with Davidson. First, I think that a sufficient and necessary analysis for action is not possible. And second, the lack of a sufficient and necessary analysis does not indicate that CTA is ontologically defective. However, my position is substantially different from Davidson's. Contra Davidson, I do not think that the lack of sufficient and necessary analysis should be explained by anomalousness of psychology. In addition, since I am defending a CTA about the nature of action, I do not think that giving up on specifying a non-deviant causal chain is a tenable option (even though I maintain that the proper specification need not be an analysis that delivers sufficient and necessary conditions for action).

### 4.1 CTA as a Reductive Explanation for Action

If the focus of defending a CTA is not on seeking the sufficient and necessary conditions for action, then what should it be? To answer this question, we should look back at the aim of CTA. As said, the CTA defended in this chapter is about the nature of action. Accordingly, it needs to achieve two aims. First, it needs to facilitate our understanding about how action can come about through the event-causal process. And second, it needs to show that action is *nothing over and above* the causal process provided in the analysis. Thus, I take CTA as a reductive explanation for action.[163] I characterize it in the following way.

> CTA as a reductive explanation: an action obtains in virtue of a specific causal structure constituted by mental states and bodily movement.

From this understanding, the key commitment of CTA is a relation of ontological dependence between action and the relevant causal structure. Particularly, this relation must be able to explain why, action, as we ordinarily understand it, can arise through this event-causal structure.[164]

---

answered independently. More importantly, one can explain the occurrence of action in an event-causal form does not follow that one can also provide an account of agent's maintaining control in an event-causal form.

[163] This idea of reductive explanation is borrowed from the discussion in philosophy of mind. It is a popular idea to take physicalism as a reductive explanation—to reductively explain mental properties (intentional and phenomenal) in terms of physical properties. See Chalmers 1996, 42–47; Kim 1998, 97–103.

[164] There are several candidates, such as realization, grounding, and logical supervenience. I take it as an open question which candidate should be endorsed as an interpretation for the ontological dependence.

Now the question is whether seeking a set of sufficient and necessary conditions for action is an essential part of this reductive explanation for action. For Bishop and many others, the answer is probably positive. As said in the first section, Bishop proposes that the aim of CTA is to show that action is constituted and realized by an event-causal process. To achieve this aim, defenders need to provide a sufficient and necessary analysis for action which applies to all possible worlds with similar ontology and causal orders. However, a reductive explanation does not always follow from a set of sufficient and necessary conditions even with modal robustness. For example, an object's having colour is sufficient and necessary for the object's having shape in all relevant possible worlds. The relation of colour and shape is concomitance, but not constitution; nor there is any conceptual connection between these two.

The crucial point of reductive explanation, I suggest, is not to look for necessary and sufficient conditions. Rather, it is to provide a functional analysis of action.[165] A functional analysis of a phenomenon/property is to specify its functional role within a system or framework, where its functional role is understood as the relations to other properties/ phenomena located in the system or framework, typically being cashed out in causal/dispositional terms. A functional analysis serves as the starting point for a reductive explanation. Only with a proper functional analysis can we see whether the target phenomenon/property can be realized and reductively explained by more basic phenomena/property. For example, to reductively explain what a gene is, we first need to provide a functional analysis of genes—that is, something which enables organisms to pass on biological traits to their offspring. And through empirical research, it turns out that this functional role is realized by specific DNA sequence, the double helix. Then we have a reductive explanation for genes. Likewise, if the aim of CTA is to explain action reductively in terms of an event-causal process, it needs to functionalize the concept of action. Only with an adequate functional analysis of action can we see whether actions can be realized by event-causal processes.

But what is an adequate functional analysis? This question seems to revive the requirement of sufficiency and necessity. For a natural suggestion is that an adequate functional analysis would still need to be a sufficient and necessary analysis. However, this is not the case. In what follows, I will first show that a functional analysis of action is by and large a functional analysis of the agent's *maintaining and exercising of control.* By providing a functional analysis of control, I will show that the requirement of sufficiency and necessity should be rejected. This is because the notion of control comes in degrees. If

[165] Bishop sometimes suggests that the concept of action should be understood as a functional concept (Bishop 1989, 143). In addition, his account is influenced by a functional conception of control from cybernetics. However, Bishop never seriously takes up the enterprise to provide a functional analysis for action (Bishop 1989, 168).

there is no way to draw a sharp line between the control and non-control, there is no way (and also no need) to provide a sufficient and necessary analysis for action.

### 4.2. The Significance of Control

If a reductive project of action requires a functional analysis of action, then a functional analysis of action is by and large a functional analysis of the agent's maintaining control. Actions by nature are the agent's maintaining control over her body and the related extension (e.g., tools & objects) to achieve some specific goals. The notion of action and the notion of control must bear a closed conceptual relationship. Any kind of action must involve at least some sort of control.[166] It is hard to imagine a case in which the agent is performing an action without control. Admittedly, an agent can be acting irrationally if he is acting out of intense emotions (e.g., furiousness) or irresistible urges. A drug addict cannot stop himself from getting drugs from the drawer even though he knows very well that the best thing to do is stop taking a drug. Aren't these cases acting without control? Surely, acting in light of good reason is a significant aspect of self-control. This drug addict does not have self-control because he cannot act following his best judgment. Nevertheless, to execute his action of getting drugs, he must possess a more basic kind of control so that he can move his limbs to open the drawer with the guidance of visual perception, etc.

Accordingly, the notion of control is crucial to the problem of causal deviance. How do we tell a causal chain is deviant or normal? A natural answer is that we do it by appealing to our intuitions. But there should be a further question about the source of these intuitions. If, as I have proposed, action by nature is exercising control, it is natural to think that these intuitions are based on our ordinary conception of control. This idea is not uncommon in the literature on causal deviance. Authors have found that the deviant cases invariably involve certain control-undermining factors. For example, Bishop (1989, 148-49) writes that: "A striking feature of deviant cases is the sheer fluke by which the intervening condition that deprives the agent of control causally contributes to the very behaviour intended." Similarly, Schlosser

[166] For a more speculative claim, regarding agency, the concept of control is prior to the concept of action. There may be some agentive control without action. In some situations, one can exercise control by not moving her body. Just imagine a soldier in ambush. There is a fly on his face which makes him very uncomfortable and he has a strong desire to get rid of it. However, he has been ordered to stay still. In such a situation, the soldier is displaying his agency and exercising his control by not doing anything. Besides, it is plausible that children develop a general capacity of controlling their bodies well before they can perform anything like an action – they practice exercising control over bodily movements, first honing that skill, then later putting those skills to work in actions. I am indebted to Jules Holroyd for the latter point. I do not have to defend the claims here. What I need is just to establish the idea that action is always associated with control.

(2007, 188) writes that "[i]n all cases of deviance, some control-undermining state or event occurs between the agent's reason states and an event produced by that agent." Aguilar (2012, 9) writes that "what all cases of deviance exhibit is precisely the undermining of the possibility for action by the undermining of agential control."[167] And finally, Shepherd (2014, 407) writes that "we can leverage an understanding of control into an account of non-deviant causation by focusing on control's possession." Given the close connection between control and non-deviant causal chains, it is plausible that our judgments about deviance or non-deviance are based on our judgments about control.

In sum, I have shown that the notion of control is essential to the discussion of causal deviance. To reductively explain action, one first need to reductively explain agent's maintaining control. The problem of causal deviance, then can be viewed as the main obstacle to reductively explaining agent's maintaining control. Given the central role of control in action, it is surprising that few action theorists have taken on the enterprise to investigate this concept.[168] Thus, in what follows, I will provide a functional analysis of the everyday concept of control. In doing so, I hope to show that our ordinary conception of control is multi-faceted and gradual, to which I refer as the *gradualist conception of control.* This conception of control further implies that a sufficient and necessary analysis for control is not possible.

## 4.3 The Gradualist Conception of Control

### 4.31 Some Qualifications

Here are some qualifications with the account that I am going to develop. First, the account is about our everyday conception of control. In some disciplines such as cybernetics, the notion of control is treated as a theoretical one. Admittedly, those theoretical discussions are helpful to illuminate our understanding of control.[169] Still, my focus is on our pre-theoretical grasp of control because it is the source of intuitions about causal deviance.

Second, what I mean by control is a two-value relation between a subject (A) and an object (B); where A and B are placeholders for anything that the ordinary judgment of control applies (it can be individuals, objects or systems). If there is control, then there is a two-value relationship that obtains between A and B—A controls B; or B is under the control of A. Much of my discussion is not restricted to the realm of agency. Our ordinary conception of control has a broader application. This is reflected in our ordinary

[167] Aguilar (2012) also refers to the control-undermining factors as 'fortuitousness'.
[168] A notable exception is Shepherd (2014).
[169] For example, Bishop's account of CTA is inspired by the study of servosystem. See Bishop (1989, 189).

language. For example, we can say that 'the room's temperature is controlled by the thermostat' and that 'the population of foxes in this area is controlled by the ecological system'.

Finally, since I am providing a functional analysis instead of a strict conceptual analysis for control, I will not try to exhaust the semantic components of control. Rather, my approach is to enumerate several factors which influence our common-sense judgments about control. These factors explain why we judge that a certain control relation is held or not held. Importantly, these factors about control are construed as functional properties. I will show that all of them can be accommodated or realized by the event-causal process depicted by the sophisticated CTA. This amount to a partial defence of the sophisticated CTA. In addition, the account of control will show that the notion of control is multi-faceted and gradual. This will eventually help defenders of CTA to liberate from the requirement of sufficiency and necessity.

**4. 32 Two necessary elements of control**

When the concept of control has application, two conditions are necessary. The first one is causality and the second is purposiveness. I will discuss these two in turn.

(1) Causality

If A controls B, then either there are causal interactions between A and B or there is a tendency/disposition of causal interaction between A and B. It is an intuitive idea that the control relation is backed up by a causal relation. If A and B are causally separated from each other, then it is hard to imagine that there is any control relation between A and B. Nevertheless, a control relation need not always to be associated with actual causal interactions. For control can obtain among objects with only tendency to causally interact. For example, Frankfurt (1978) has devised a scenario in which a driver's automobile is coasting downhill in virtue of gravitation. Even though the driver does not actually intervene or adjust the automobile, the automobile is still under control. This is because the driver can interfere if he wants to. Or suppose a thermostat controls the indoor temperature. If the current temperature fits with the pre-set value, then the heating or cooling mechanisms do not need to operate at all.

The factor of causation fits well with the CTA. Proponents of the CTA can admit that an action is an agent's exercising control over her bodily movement. They can also admit that to exercise control, the agent's mental states must bear some causal relations to her body.

(2) Purposiveness

The concept of control only applies to goal-directed phenomena. Specifically, if A controls B, then A must possess specific goals with respect to B. Aimless causal interactions do not constitute control relations. For example, within the solar system, the sun and the earth are causally interacting with each other through gravitation. However, there is no control relation between the sun and the earth. This suggests that apart from causality, purposiveness is another necessary condition for control. A goal is a pre-set function of a system. It can be obtained in different contexts. Intellectual beings have goals because of intentional mental states. A machine can have a goal because of the purpose of the designer (the thermostat is designed to maintain heat). A biological organ can have a goal because of evolution (e.g., the human hearts possess the function of pumping blood due to natural selection).

This factor can also be accommodated with the CTA. According to the CTA, if the agent is performing an intentional action, the action must be caused by specific mental states such as desires and intentions which represent the agent's goal or purpose. Now I would delineate some factors that would interfere with our judgement of control.

**4.33 Three Criteria for Control Assessment**

I have outlined two necessary components for a control relation to be held, namely, causality and purposiveness. However, these two are still not sufficient to establish a control relation. Intuitively, control is a normative notion. We can say A exercises high-level control over B. Equally, we can say A controls B badly. Thus, the concept of control must have certain normative dimensions. There should be some criteria to assess whether the control relation obtains and to what extent a control relation obtains.

(3) Accuracy

Since control aims at achieving goals, the direct way to assess whether A controls B successfully is to see how accurate the control outcome matches the original goal. That is to say, if A controls B, then there should be a match between A's goal and the actual outcome through the causal manipulation of B. For example, a thermostat is set to keep the room temperature at 16 degrees. If the room temperature is maintained as the target number due to the thermostat, then the thermostat controls the temperature well. Note that accuracy is also a notion that comes in degrees. In our ordinary judgement about control, we usually do not require a perfect match between the goal and the outcome. If the thermostat at issue can

only keep the temperature varying from 14 degrees to 16 degrees, then we would still judge that the thermostat maintains control, though not to a high extent. Very roughly, if A's manipulation of B achieves the goal more accurately, then A controls B to a higher degree.

(4) Reliability

Accuracy alone is not the only criterion we use to evaluate control. Suppose a novice dart player happens to hit the bull's eye in his first try. Even though he accomplishes the task with accuracy, we tend to think this is due to luck rather than control. To judge that there is genuine control, we not only require accuracy, we also require accuracy can occur repeatedly in *similar* actual or hypothetical situations because we hope to exclude the possibilities of attaining accuracy by accidental factors. There should be an emphasis on the reliability of the causal mechanism from which the subject is exercising control. When holding fixed the measure or mechanism through which A manipulates or interacts with B, reliability is the average accuracy of the match between the outcome and the goal of A. The factor of reliability serves as a supplement to the factor of accuracy. This is because accuracy only describes the match between the outcome and the goal at one time, while reliability reflects the average accuracy over time. Reliability is a factor that comes in degrees. For example, a dart player has a 50% chance to hit the bull's eye, another player has 60%. We can say that the latter is more reliable in hitting the target.

(5) Flexibility

Reliability is a measurement of the control performance in similar situations, while a controlling subject or system usually comes across different situations and obstacles when exercising controls. Thus, we also need a criterion to evaluate whether the controlling subject can manage a wide range of different situations. Flexibility is the ability of the controlling subject to adjust the goal and causally interact with the object correspondingly. It consists of two parts: firstly, the controlling subject is able to regulate goals in responding to the properties of the situations. This requires the subject to exchange information with the object and the environment; and secondly, the controlling subject is able to causally interact with the object in accordance with the goal's variation. If A controls B, then A's goal is responsive to the changing situations; meanwhile, A's causal interaction with B is responsive to the changing goal of A. For example, a thermostat is set to keep the room temperature as 15 degrees. The working mode of the thermostat depends on the room's instant temperature. The thermostat keeps detecting the room temperature. When the room temperature is higher than the target temperature, then the current goal of the

thermostat will be lowering the temperature. Correspondingly, the cooling mechanism begins to work; when the room temperature is lower than the target one, the heating mechanism begins to work. In this simple control system, the control relation holds because the thermostat can receive information regarding the temperature and adjust its working mode in response to the new mode.

Note that the notion of flexibility is a notion that itself comes in degrees. A controlling subject can receive the information input in different subtleties and react to it in different sensitivities. Compare two different thermostats. One can only detect a subtle temperature change and react to it sensitively, say, it can be set to maintain temperature as accurately as to the first place of decimal (e.g., 15.0 degree or 15.1 degrees). The other only react to the temperature more coarsely. It can be set to maintain temperature as accurately as to one digit (e.g., 15 degrees or 16 degrees). We can say that the first thermostat exercises more flexible control than the second one over the temperature.

**Summary: The Gradualist Conception of Control**

Based on the discussions above, I have two conclusions. First, the notion of control is multi-faceted. There are different kinds of considerations that may have an impact on our judgement about control. These considerations include causation, purposiveness, flexibility, accuracy, reliability and exclusivity.

Second and more importantly, our notion of control comes in degrees. I call it the gradualist conception of control. This means that our judgement of control is not a judgement about all or nothing. Instead, our judgement of control may have many values: high degrees of control, low degrees of control, or just high degrees of control in some of the respects but not others etc. This is because, first, some of the factors which influence the judgement of control do come in degrees. For example, a control with higher degrees of flexibility may accordingly be in higher degrees in itself. And second, the judgement of control is co-determined by different factors. Different combinations of factors with different significances may generate different judgements about control. For example, a control with significant flexibility but little reliability may be judged differently from the one with significant accuracy but little reliability.

### 4.4 The Gradualist Account of Control and the Causal Theory of Action

I have provided several aspects of our ordinary conception of control. These aspects together serve as a functional analysis of control. Now the question is whether these aspects can be accommodated by an event-causal process? My answer is that an analysis from the sophisticated CTA can do justice to all these

aspects. Recall that according to the sophisticated CTA, action arises from an ongoing causal interaction between the agent's control system and the bodily movements; where the control system is constituted by various mental states. Thus, action can be taken as a control relation between the control system and the bodily movements.

Let's first look at the two essential components of control. Since CTA construes action as an event-causal process, thus the aspect of causality is inherently taken into account. In addition, the causal process is goal-directed. This is because the control system is constituted by mental states, where goals are represented.[170]

What about the other gradable dimensions of control? The dimension of accuracy has already been accommodated by the traditional version of CTA. Accuracy can be taken as the match between the action-outcome and the goal represented by the motivational mental states. When a proponent of CTA claims that an action is a bodily movement caused by specific motivational mental states, he implicitly presumes that this bodily movement is what that the agent intends to occur. Otherwise, it is just a failure of action performance rather than an action.[171]

Reliability is a little tricky for it is a modal property and cannot be directly reflected in a causal route from the motivational mental states to bodily movements. This is related to a difficulty with the CTA that the exercising of ability cannot be accommodated in the causal process.[172] To illustrate, compare the case of a novice dart player hitting the bull's eye and the case of an experienced dart player hitting the bull's eye. Both actions can be cashed out in the same causal process—an intention to hit the bull's eye cause the matching bodily movement. However, there must be some differences in these two causal processes for one involves exercising an ability and the other does not. Fortunately, this problem can be solved by the sophisticated CTA. Recall that the causal process depicted by the sophisticated CTA involves multiple levels. On the personal level, a skilful action and a non-skilful action may be of the same causal-type, while on the sub-personal levels, they are actually underpinned by different causal mechanisms. For example, actions of different skill levels are underpinned by causal processes involve different motor schemata, which are fine-grained representations of more basic action units.[173]

The factor of flexibility can be captured by the CTA from several respects. First, whenever an agent changes her mental states, this change can be reflected in her change of behaviour such that she may

[170] Of course, how mental states obtain intentionality is a difficult problem, which I cannot address here.
[171] For example, see Shepherd (2014).
[172] This difficulty is raised by Hornsby (2004; 2008).
[173] See Jeannerod (1997). Aguilar (2012) proposes another approach according to which the reliability of the causal mechanism underlying the action can be cashed out in terms of "action repertoire".

change or discontinue the ongoing behaviour. Second, it can be captured by the sensitivity condition: if the content of the intention changes slightly, then the bodily motion should be changed correspondingly. Third, it can be captured by the feedback loops: when an agent is acting, she is adjusting her bodily motion in accordance with the feedback information.

### 4.5 The Gradualist Conception of Control and the Problem of Causal Deviance

Why does the gradualist conception of control have anything to do with the problem of causal deviance? Recall that at the beginning of this section I provide several suggestions about where the intuitions about causal deviance come from. My suggestion is that the intuitions are rooted in our pre-theoretic grasp of control. If the notion of control comes in degrees, it implies that there is no definite judgement about deviance/non-deviance. In effect, there is no way to provide a set of sufficient and necessary conditions for non-deviant causal chains.

As mentioned, many philosophers hold that a successful solution to the problem of causal deviance must provide sufficient and necessary conditions for a non-deviant causal chain. However, if our notion of control is multi-facet and comes in degrees, then the seeking for the adequate analysis of non-deviant causal chains is doomed to failure. For different actions may require different amounts of control or involve different aspects of control. Whenever a revised analysis of CTA is provided in an attempt to exclude all deviant causal chains, there will be counter-examples indicating that either the revised theory is too strong or too weak or both. This claim is vindicated by the current states of the debate: defenders of CTA have established different conditions for non-deviant causal chains to circumvent the cases of causal deviance while the new conditions seem not to handle the new contrived cases of causal deviance such as the case of double-agent deviance and the case of a sensitive fluke (e.g., the case by O'Brien mentioned in 2.3.2).

According to my account, the impossibility of providing the sufficient and necessary conditions does not entail that we should abandon the CTA or the reductive project of action. As discussed, the failure to provide sufficient and necessary conditions does not result from the event-causation framework, but from the fact that our judgements about control are multi-faceted and comes in degrees.

## 4.6 Objections and Replies

### 4.6.1 The Objection of Threshold

One objection to my argument is that even if the judgement of control comes in degrees, there is still some threshold requirement of control for a bodily motion to become genuine action. It then follows that in principle there can be sufficient and necessary conditions for control if those thresholds are made explicit. I call this objection the objection of threshold. Admittedly, there must be some minimum control requirement for each action to be counted as an action. However, there are still two reasons to reject that a sharp line can be drawn between the deviant-causal chains and the non-deviant causal chains.

Firstly, 'action' is a term used in ordinary language where vagueness is common. In ordinary language, there seem to be no clear criteria to determine how much hair loss would be counted as bald. The vagueness of language makes borderline cases always possible even in typical sufficient and necessary analysis, such as the one that a bachelor is an unmarried man. What to do with, say, a single 15 years old boy? [174] If 'action' is a term employed in ordinary language, borderline cases can be easily found and we will be reluctant to count them as actions or non-actions. Of course, we can always draw a line at will. But this line seems not to represent anything substantial to distinguish action from non-action.[175]

Secondly, I think it is untenable to invoke a single universal criteria-set of control for all the action-types in all the performing contexts. There are some action-types or contexts which require more expertise or dexterity, say in the context of professional sports games. And some other contexts are requiring less. Imagine a patient suffered a severe injury in his arm. After the operation, a doctor asks the patient to raise his arm to test the effect of the operation. In this context, even if the patient cannot exercise a high level of control, his raising hand is an action because it achieves the aim in the context. If we set a too high a standard, then it may exclude some actions we do want to include. If we set the standard too low, we may then include some bodily motions that we do not want to. In different situations or for different action-types, we expect different criteria for maintaining control in the action, this makes it difficult to set up universal criteria for control in action.

[174] This example is borrowed from Cleland and Chyba (2010) where they discuss the definition of life.
[175] Cf. Mele (1992, 6)Some of our concepts are simply fuzzy around the edges; and stipulative tidying up can be more trouble than its worth."

### 4.6.2 The Objection of Conflating Vagueness with No Sufficiency and Necessity

Another objection is that my argument at best shows that the concept of control is vague. But this still does not explain why a sufficient and necessary analysis for action cannot be obtained. For vagueness is not the same as lacking a set of sufficient and necessary conditions. Specifically, if an analysis is vague, it entails the existence of borderline cases. By comparison, if a concept lacks sufficient and necessary conditions, it entails that for any possible analysis of that concept, there are always counterexamples.

My reply to this objection is twofold. First, in my account, it is not just vagueness of the notion of control that precludes us from providing a sufficient and necessary analysis for action. Rather, it is vagueness plus multifariousness of the notion of control that makes it difficult to provide a sufficient and necessary analysis. Different action-types would probably highlight different dimensions of control and downplay others. Thus, it would be different to have a sufficient and necessary analysis which applies to all action-types.[176]

Second, I admit that vagueness will not preclude a sufficient and necessary analysis. Nevertheless, vagueness will preclude an *informative* sufficient and necessary analysis. Without informativeness, a sufficient and necessary analysis comes cheap. Here is one: an action is a bodily movement caused by specific motivating mental states in a non-deviant way. Nevertheless, no one will find this analysis satisfactory. For we still need to know what makes a causal chain deviant or non-deviant. My contention, precisely, is that there is no informative sufficient and necessary analysis for action which universally determines a case to be actional or non-actional. If we want to be informative enough, we have to set each bar of control dimension conclusively. But as argued, this cannot be achieved.

Admittedly, a borderline case is different from a counterexample. A borderline case is a case which does not clearly satisfy the postulated analysis of a concept; in effect, a borderline case is not clearly an instance of the concept. By comparison, a counterexample is a case which clearly satisfies the postulated analysis of a concept but intuitively it is not an instance of the concept. However, as the analysis of the concept is not informative enough, the distinction between borderline cases and counterexamples sometimes fades out. Consider the analysis that a bachelor is an unmarried man. What about a separated

[176] Compare the notion of family resemblance introduced by Wittgenstein (1952/2009). The example Wittgenstein talks about is the concept of game. There are bunch of typical properties which are associated with the concept of game (e.g., being competitive, being played on a board, being played with cards, being played with a ball, and so on). These properties in the bunch are neither sufficient nor necessary. For every particular game-type instantiates some of the properties in the bunch while omit the others. I am not proposing that the concept of action is a concept of family resemblance. For each action-type must have all the dimensions of control listed above. However, there is still one similarity here. That is, different action-types may involve different degrees in each control dimension. And for the similar reason, we cannot have a sufficient and necessary analysis of action.

man? Is it a counterexample or a borderline case of the analysis? This question is not easy to answer because 'being married' is not informative enough to include or exclude 'being separated'.

My argumentation predicts that the counterexamples to the sophisticated CTA analysis can be better regarded as borderline cases. I think this prediction matches the dialectical state of the literature. As the causal deviance cases become more and more convoluted in order to satisfy the conditions in a CTA analysis, our intuitions become less and less decisive. Recall the case of deviant heteromesial cases which involves feedback loops signals that go back to both the agent and the neuroscientist. This case is supposed to serve as a counterexample to the sufficiency of CTA. However, since this case is complicated, I do not think that our verdict on this case is decisively deviant and non-actional. In addition, since in this case the matching behaviour of the agent entirely depends on the neuroscientist, it is not clear that the bodily motion of the agent can satisfy the dimension of reliability, which is essential to our judgments of control. Therefore, unlike classical cases of causal deviances (such as Climber and Robber) which definitely serve as counterexamples, the deviant herteromesial cases involving feedback loops should be better regarded as borderline cases of action.

**Conclusion**

In this chapter, I have deflated the assumption that a successful solution to the problem of causal deviance hinges on a sufficient and necessary analysis for action. Specifically, I develop an account of control to which I refer as the gradualist account of control. I have shown that the notion of control involves different gradual dimensions and each dimension can be accommodated by the sophisticated CTA developed by Peacocke and Bishop. In addition, based on the gradualist account of control, I argue that the lack of sufficient and necessary analysis should not be explained by the inadequacy of event-causal framework, but explained by the fact that our judgments of control are multifaceted and comes in degrees. In effect, there is no sharp line to be drawn between action and non-action.

There is another objection to CTA. That is, even if a CTA can show that the exercising control in an action can be reductively explained within an event-causal framework, the crucial point is to reductively explain the agent's exercising control. It is the roles of the agent that are omitted in an event-causal framework. I will tackle this objection in the next chapter.

## Chapter 5: Disappearing Agency and Two Intuitions about Action

**Abstract：** In this chapter, I tackle the problem of disappearing agency which plagues the causal theory of action. I first enumerate several possible interpretations for the problem, and then I point out certain conceptual issues related to the problem which are the most pressing. Specifically, under one interpretation, the problem is that the event-causal framework employed by the causal theory of action cannot accommodate two intuitions about action, namely, Agent-Participation (that agent participates in her action) and Anti-Reduction (that agency as active phenomena cannot arise from merely event-causal interactions). To solve this problem, I develop a positive account, namely the structural account, to show that agent's roles in her action can be fulfilled by the causing of her mental events as well as her psychological structure. I also provide arguments to explain away the intuition of Anti-Reduction.

### 0. Introduction

The problem of disappearing agency is regarded by many philosophers as one of the most severe challenges to the causal theory of action (CTA henceforth). This problem is highlighted and has become well known due to Velleman's 'What happens when someone acts' published in 1992. [177] In this article, Velleman raises specific concerns to the standard CTA (which he refers to as the 'standard story'). According to the standard CTA, action is a bodily movement caused by the agent's motivating mental states (e.g., desires, beliefs, intentions) in a non-deviant way. Velleman's key claim is that the description of action delivered by the standard CTA makes the agent disappear. In a widely quoted passage, Velleman writes:

> I think that the standard story is flawed in several respects. The flaw … is that the story fails to include an agent— or, more precisely, fails to cast the agent in his proper role. In this story, reasons cause an intention, and an intention causes bodily movements, but nobody—that is, no person—does anything. Psychological and physiological events take place inside a person, but the person serves merely as the arena for these events: he takes no active part. (Velleman 1992, 461)

[177] Similar concerns were adumbrated before Velleman's article, e.g., Taylors (1966), Chisholm (1978) and Nagel (1986)

At a first approximation, the criticism is that the agent or the roles of the agent cannot be accommodated in the description of action proffered by standard CTA. In effect, the standard CTA fails to capture an essential idea of action—that is, action is the agent's doing something. Philosophers have different reactions to this problem. Some philosophers including Velleman take it as only a challenge to traditional versions of CTA (particularly the standard CTA). Some philosophers hold that this problem is fatal to all versions of CTA (regardless the traditional versions or the subsequent improved ones). There are also some philosophers who dismiss this problem as a non-serious issue. For example, Schlosser, a defender of CTA, once complains:

> [Proponents of the problem of disappearing agency] have not produced a single argument to support their case, and they have certainly not identified a philosophical problem. Their case is entirely based on intuition, and in some cases on mere metaphor and rhetoric. (Schlosser 2010, 22).

To some extent, I am sympathetic to Schlosser's point that the nature of the problem has seldom been explicitly formulated, but I do think that there is a way to formulate it and it is worthy of serious consideration. This is the task of the present (and next) chapter: to explicate, and defend CTA against, the problem of disappearing agency.

The different reactions from philosophers suggest that there are different philosophical concerns under the name of the 'problem of disappearing agency'. To tackle the problem of disappearing agency, we should first carefully tease out these concerns and then assess the potential arguments backing these concerns. In the preceding chapters, I have characterized CTA as a program of reductive explanation. That is, actions can be reductively explained by the causal interactions among mental states/events and bodily movements. The implication of CTA is that action has to take place within an event-causal framework. With this clarification, we can interpret the quoted text from Velleman in two different ways:

The first gloss is that the psychological description (typically, with beliefs, desires, intentions causing the bodily motions) provided by standard CTA is too simplified to accommodate the roles of the agent during an action. Accordingly, the standard CTA is incomplete but not fundamentally wrong. It can be fixed by incorporating a more detailed psychological story which can do justice to the role of agents. Importantly, this detailed psychological story can be told within the event-causal framework. Many proponents of CTA are working on the problem under this reading.[178]

[178] E.g., Velleman (1992), Bratman (2001), Enc (2003), Mele (2003).

The second gloss is that the ontological framework, namely, the event-causal framework, fails to capture the phenomena of agency. According to this reading, no matter what sophisticated psychological process is added to the story, CTA is not able to capture the essential aspects of action. Thus, under this reading, the problem of disappearing agency, by nature, urges philosophers to abandon the event-causal framework, which boils down to abandoning the CTA project entirely. [179] Many opponents of CTA are taking up the problem under this reading.[180]

Since one of my main concerns in the whole thesis is whether our ordinary conception of agency can survive formulation within the event-causal framework, I will primarily focus only on the second gloss of the problem.[181] Specifically, the problem of agency under the second reading can be formulated as the incompatibility of the event-causal framework and the phenomena of action and agency. I diagnose that this impression of incompatibility is rooted in certain intuitions about action and various kinds of phenomenology of agency. I will focus on the intuitions in this chapter and leave the phenomenology to the next chapter.

There are two intuitions about action which together motivates the worry that agency disappears in the event-causal framework. I call these two intuitions *Agent-Participation* and *Anti-Reduction* respectively. Below I will articulate these two intuitions and assess the related arguments.

**1. Reductive Explanation and Agent-Participation**

As mentioned, CTA aims to reductively explain action in terms of event-causation. However, there are some features of action which seem to resist such kind of reductive explanation, namely, Control-Maintenance and Agent-Participation. In the last chapter, I investigate the possibility of reductively explaining Control-Maintenance. Below I will first cash out the idea of Agent-Participation. And then I will review certain possible arguments for the claim that Agent-Participation is an obstacle to reductively

[179] Pereboom (2014) also uses the notions of "disappearing agent objection" to denote a criticism to the indeterministic event-causation account. Specifically, Pereboom holds that the event-causal libertarianism cannot accommodate the role of agent in settling a specific outcome in an indeterministic causal chain and in effect makes the agent disappear. See also Clarke (2017). This reading of disappearing agency does not fall onto either my first gloss and second gloss. Since Pereboom's objection is specific to the debates in free will while my focus in this chapter is on action theory, I will set this reading aside.
[180] E.g., Hornsby (2008), Steward (2012).
[181] Admittedly, a complete solution to the second gloss of the problem has to address the first gloss of the problem.

explaining action in in terms of event-causation. I will deflate these arguments and will show that Agent-Participation obtains in an event-causal framework.

Agent-Participation is the intuition that the agent must participate in producing her action.[182] This intuition is widely shared and entrenched in our ordinary conception of agency. Suppose Jack is flinching because of intense pain. Although his flinching is his bodily movement, we would not take it as the action done by Jack. And the reason is probably that it is not Jack who initiates his own flinching. Rather, the movement is brought about by the pain, which is just a mental state that befalls on him. Or consider the scenario of Mike who can act just like us (he chooses a snack, decides to write a paper, etc). One day it is discovered that all of Mike's actions are produced and controlled by a remote evil neuro-scientist through a device implanted in his brain. Then, we no longer regard the bodily motions of Mike as genuine actions. Even worse, we no longer regard Mike as a genuine agent who can act, for he has no role to play in his action.

Since Agent-Participation is one of the most important characteristics of action, to reductively explain action, we need to reductively explain Agent-Participation. But reductively explaining Agent-Participation within an event-causal framework is not easy. There are two concerns. The first is about the order of causation. Agent-Participation seems to entail a substance-causation account of action, which is incompatible with the event-causal framework. The second concern is related to personal identity. By reductively explaining Agent-Participation in terms of event-causation, CTA seems to imply a problematic view of personal identity, namely, the self is identical to a bunch of mental states. These two concerns together contribute to the conviction that Agent-Participation cannot be realized within an event-causal framework, which in effect leads to the worry that the event-causal account of action makes the agent disappear. In what follows, I will review these two concerns in order.

[182] I will take this intuition different from the claim of ultimacy according to which the agent has to be the *ultimate source* of her action to bear real responsibility for that action. The latter is committed by many incompatibilists in the debate on free will and moral responsibility (e.g., Kane (1996), Pereboom (2001), Strawson (1994)). Albeit I am open to the idea that these two claims may be interrelated, I hold that these two intuitions are different. First, Agent-Participation is weaker—it does not necessarily require the agent to be the ultimate source of actions as many incompatibilists require; and second, the Agent-Participation is feature of action as we ordinary understand it while ultimacy is supposed to be a condition for moral responsibility.

### 1.2 Event-Causation and Agent/Substance-Causation

Many opponents of CTA think that reductively explaining Agent-Participation within an event-causal framework is doomed to failure. Here is the argument. The agent participates in her action by directly bring about her action. This idea entails a specific kind of causal relation, namely, agent/substance-causation. In this kind of causal relation, the agent figures as an irreducible substance in the explanation of action. This causal relation does not exist within the event-causal framework committed by CTA. Here is a quote from Charles Taylor, who is a famous defender of the agent-causation view:

> In describing anything as an act there must be an essential reference to an agent as the performer or author of that act, not merely in order to know whose act it is, but in order even to know that it is an act at all. ... Another perfectly natural way of expressing this notion ... is to say that, in acting, I make something happen, I cause it, or bring it about. (Taylor 1966, 109–11)

To assess the challenge, we should discern event-causation and substance causation. First, substances and events belong to different ontological categories. In ordinary language, we have different predicates for substance and event: we say that a substance exists, and an event occurs. Besides, substance and event bear different relations to space and time. Substance, especially ordinary objects such as tables, billiard balls and water, are persisting entities locating at specific spatial points—they remain relatively stable and identifiable within a certain period of time. By comparison, events are temporal particulars and are more often taken as changes rather than persistence. [183]

Those ontological differences between substance and event explain why event-causation and substance-causation must be different frameworks. In the event-causal framework, the relata of causal relations are exclusively events. By comparison, in the substance-causal framework, the cause is a substance, and the effect is an event. Even for those who support the substance-causal framework, there is a disagreement about the scope of this framework. Some hold that causation is universally substance-causation. Others hold that only in the realm of agency is there substance-causation; and in the realm outside agency, the causal relations are event-causation. I will set aside this minor difference, and use the term 'agent-causation' instead of 'substance-causation' to denote the claim that action is caused by the agent as an irreducible substance. Nevertheless, we should bear in mind that agent-causation is a sub-class of substance-causation for there are many other beings belonging to the category of substance apart from

[183] The points are extracted from Casati and Varzi (2020). You can find more detailed discussion and a complete survey of the literature in their article.

human agents. A case from the literature may help to illustrate the difference between event-causation and agent-causation in action theories. Here is a quote from Chisholm, who defended agent-causation view in his earlier writing:

> I shall borrow a pair of medieval terms, using them, perhaps, in a way that is slightly different from that for which they were originally intended. I shall say that when one event or state of affairs (or set of events or states of affairs) causes some other event or state of affairs, then we have an instance of transeunt causation. And I shall say that when an agent, as distinguished from an event, causes an event or state of affairs, then we have an instance of immanent causation. (Chisholm 1964, 494)

Though the terminology used by Chisholm is different from the one used currently, the quote represents a typical view of agent-causation. By transeunt causation, Chisholm means event-causation while by immanent causation, he means agent-causation. In this account, agent-causation and event-causation are ontologically different. For example, if an agent raises his hand intentionally, according to the agent-causal account, this intentional action *cannot* be reduced to event-causal interactions, such as one's desires, beliefs, intentions or the like causing the hand to rise. Rather, this action is agent-caused in the sense that the bodily movement is caused by the agent as a substance which is *ontologically different* from, say, his mental states or mental events.

Now return to the intuition of Agent-Participation. It is tempting to think that the agent participates in her action through directly causing her action as an irreducible substance. This idea motivates many opponents of CTA to believe that action has to be agent-caused if Agent-Participation obtains. But this idea is not well-founded. For the intuition of Agent-Participation seems to be an intuition of the common-sense level and that it is not fine-grained enough to determine nuanced metaphysical issues. Merely from the intuition of Agent-Participation, it is difficult to reach the conclusion that action is caused by the agent as an irreducible substance. The intuition of Agent-Participation seems not to involve much ontological commitment, as expected. Two more points can substantiate my diagnosis.

Firstly, in our ordinary discourse of causation, we seem to use substance-causation talk and event-causation talk interchangeably to report the *same* causal phenomena. For example, we can say that a *ball* causes the break of the window, which is a typical claim of substance-causation. But we find it equally natural to say that the *flying* of the ball causes the break of the window, which is a claim of event-causation. In our ordinary discourse, we do not make a sharp distinction between substance-causation and event-causation. This may further suggest that our common-sense conception of causal phenomena is not

fine-grained enough to determine the metaphysical disagreement about causal frameworks. This lesson applies to the case of agency as well. As said, agent-causation is usually regarded as a special case of substance-causation (e.g., the flying ball cause the window to break). Although we do employ agent-causation terms to describe agency and action, it does not follow that the use of these terms involves substantial ontological implications.

Secondly, Agent-Participation, as an everyday intuition, does not involve much ontological commitment about the notion of agent. Our ordinary conception of agent, or other related concepts, such as person, or the self, is neutral on the metaphysical questions such as whether the agent is a bunch of mental states or a non-reducible substance. This leaves much room for proponents of CTA to reconstruct the idea of Agent-Participation in terms of causation of agent-involving events. The proponents of event-causation accounts of action can claim that their theories are also abiding by Agent-Participation as long as they argue that the event-causation account can accommodate the intuition of Agent-Participation. For example, Robert Kane, who is also a proponent of event-causation account of action, writes that: "[d]oing without agent-causation in the non-occurrent sense does not mean denying agent-causation in the ordinary sense that agents act, bring things about, produce things, make choices, form their own characters and motives, and so on." [184] What Kane proposes here is that one can talk like an agent-causalist in common-sense term, whilst being an event-causalist at the ontological level. To show that CTA can accommodate Agent-Participation, what proponents need to do is to argue that agent-causation can be realized by and reduced into event-causation. For example, Clarke (2017) proposes a schema to reduce agent-causation into event-causation as follows:

> Substance s caused event e2 just in case there was some event, e1, such that e1 involved s and e1 caused e2. (Clarke 2017, 2)[185]

With this schema, a proponent of CTA can then analyse the common-sense agent-causation in terms of evet-causation. For example, an agent (s) raises her hand up (e2) if and only if the relevant mental states of the agent (e1) involving the agent (s) and cause her hand rising (e2) in the appropriate way.

[184] Kane (1996, 123)

[185] See also Ekstrom (2000, 114). Lowe (2008, chap. 6) provides a similar schema for event-causalists to analyse substance-causation as Clarke, though Lowe does not not think such reduction is tenable because of certain philosophical concerns. Especially, Lowe thinks that only substances have causal efficacy while events are powerless. I set aside these metaphysical issues.

### 1.3 Self-Determination and Personal Identity

I have shown that Agent-Participation as a common-sense intuition does not necessarily commit to irreducible agent-causation. For our common-sense intuition is neutral on metaphysical issues such as the fundamental order of causation and the nature of the self. However, there are still philosophical concerns with explaining Agent-Participation in terms of event-causation. In particular, the reduction may commit to a view of personal identity which cannot withstand philosophical scrutiny. Recall the intuition of Agent-Participation—that action has to be something that the agent participates in bringing about. To reductively explain Agent-Participation, defenders of CTA have the burden to show that the roles of the agent in her action are all realised in event-causation (cf. Clarke's account, canvassed above). This move seems not only to reduce action into causal interactions among events, but also commits to reducing the agent into events or set of events. Thus, some may suspect that this move would make the agent disappear or at least distort our conception of the self.

Before presenting the challenge from personal identity, I should first deflate a naïve understanding of the criticism, which can be formulated in the following way: according to standard CTA, an action is just a bodily motion caused by certain motivating mental states in a non-deviant way; since this description does not directly refer to the agent, thus, the agent disappears in the description from standard CTA. However, this reading is unfair to defenders of CTA, as Velleman rebuts it in his article: "[c]omplaining that the agent's participation in his action isn't mentioned in the story is, in their view, like complaining that a cake isn't listed in its own recipe".[186] The problem with this naïve understanding is that it conflates elimination and reduction. To eliminate an entity/property means to deny the existence of such entity/property in our ontology; to provide a reductive analysis of an entity/property is to reductively explain the entity/property in terms of more basic components while still admitting the existence of the entity/property. For example, saying that witches do not exist is to provide an eliminative theory about witches. Saying that temperature is the movement of molecules is providing reductive theory about temperature. The description from CTA does not mention 'agent' in its story because CTA aims to explain agency by providing a reductive account of action.

Here is a more charitable way to read the criticism. It is problematic to reduce the agent or the self into certain agent involving mental states. Particularly, in the story provided by CTA, it is the set of relevant mental states that cause the bodily movement. In this story, to say that the action is caused by the agent is to say that the action is caused by the relevant mental states of the agent. This seems to leave us the

[186] Velleman (1992, 462)

consequence that the agent has to be reduced to the mental states. However, to equate the agent with a certain set of mental states would surely lead to unpalatable consequences.

This problem has recently been highlighted by Christophe Franklin (2016; 2018, chap.7), who argues that the reductive project of CTA commits to a problematic view of personal identity. Franklin's argument begins with the observation that the agent plays certain roles in her action, which are more than simply bringing about the action when she is acting. Franklin refers to these roles as self-determination. Specifically, the agent not only causally initiates the action, she also adjudicates between her various motivations and makes a decision based on that adjudication. According to Franklin, self-determination can be highlighted in a special kind of experience in which we as agent can stand back behind our mental states and intervene with those mental states. He writes:

> This experience is most vivid in the case of motivational conflict. Consider the classic case of conflict between duty and desire, such as when I know I should be more attentive to my children but am exhausted. Duty pulls one way, desire another. Suppose that I make an effort to turn my attention to my children and succeed in this endeavor. In this case, it does not seem that my decision was merely a function of my desires and beliefs. These attitudes were in conflict, after all, and I myself had to decide how to resolve the conflict, or so it seems. (Franklin 2018, 182)

I refer to this experience as detachment—that the agent is detached from his beliefs and desires, and exerting an additional causal influence on these mental states or events. The important message delivered by this phenomenology is that the agent's roles in action are not only simply causing the action; rather, the agent also stands behind the mental states which are causally responsible for the action, and adjudicate the conflict of these mental states. Franklin summarizes these roles as self-determination.[187] For Franklin, the event-causation account is untenable for it leaves all the causal work to be done by the mental states. Consequently, the agent's self-determining causal influence on his action is only realized by the causation of mental states or events. This will lead to a problematic view of personal identity. That is, the self is identical to a bunch of mental events or states. He refers to this argument as the *It Ain't Me Argument*, which can be formulated as follows:

[187] Even though Franklin's argument is motivated by this phenomenological consideration which by itself seems to suggest an agent-causation view of action, most parts of his reasoning seems to be independent of the content of this phenomenology.

(P1) If an agent S's causal contribution to his decision $\varphi$ is exhausted by the causal contribution of some bundle of states and events (e.g., his reasons), then S self-determines $\varphi$ only if S is identical to (some members of) this bundle of states and events

(P2) An agent is not identical to any state or event or to any bundle of states and events.

(P3) Therefore, if an agent S's causal contribution to his decision $\varphi$ is exhausted by the causal contribution of some bundle of states and events (e.g., his reasons), then S does not self-determine $\varphi$.[188]

If this argument is sound, it will lead to the conclusion that the picture delivered by CTA contains no self-determination. Since a key premise from this argument is that the self is not identical to mental states (P2) and in effect cannot be reductively analysed within an event-causation framework a further implication of this argument is that to accommodate self-determination, the agent must be a substance which is ontologically irreducible to the mental states or mental events.

Franklin's defence of P2 is compelling. He argues that reducing the agent into a bunch of mental states cannot meet two necessary conditions at the same time—the modal adequacy and the extensional adequacy.[189] The condition of modal adequacy requires that the bunch of mental states are essential to the agent's identity—any changes of mental states in the bunch will result in a substantial change of the identity of the agent. This means that the bunch of mental states have to be special and narrow to exclude normal mental states (e.g., normal beliefs or desires), otherwise the mental states are not essential enough to constitute the agent's identity. The condition of extensional adequacy requires that for every self-determined action, at least one member of the bunch has to figure in the causal explanation of that self-determined action. This means that the bunch of mental states has to be broad enough to cover all self-determined action. Now comes the dilemma: the bunch of mental states cannot be too broad otherwise it violates the modal adequacy; nor can it be too narrow, otherwise it violates the extensional adequacy.

However, many defenders of CTA will probably find P1 unfair. For they will not endorse the claim the agent is to be equated with the mental states even if the roles of the agent are exhausted by the causing of the event. Velleman provides two examples to illustrate this point. The agent's digesting food is exhausted by his digesting system; this does not follow that the agent is identical to his operation of the

[188] This argument is from Franklin (2018, chapt.7). Franklin (2016) provides another argument which is characterized in a slightly different way.
[189] Franklin 2018, 190

digesting system. Similarly, the agent's fighting against diseases is exhausted by his immune system, this does not follow that the agent is identical to his immune system.[190] By the same analogy, from the fact that the agent's performing intentional action is exhausted by the causing of mental events, it does not follow that the agent is identical to the mental events. Rather than proposing that the agent is identical to a bunch of mental states, many defenders of CTA argue that the agent is represented by certain mental states. Accordingly, the agent's role is played by those proxy mental states. Call this the *proxy account.*

Nevertheless, Franklin's argument can be reconstructed in a more charitable and defensible way to cover the proxy account:

> (P1*) If an agent S's causal contribution to his decision $\varphi$ is exhausted by the causal contribution of some bundle of states and events (e.g., his reasons), then S self-determines $\varphi$ only if
>
> i) S is identical to (some members of) this bundle of states and events;
>
> or that
>
> ii) S is represented by particular mental states or events in the sense that the role of S in the action (namely, self-determination) is played by those particular mental states or events.
>
> (P2*) An agent is not identical to any state or event or to any bundle of sates and events; nor is the agent represented by his mental states in the sense that his roles are played by those mental states.
>
> (P3*) Therefore, if an agent S's causal contribution to his decision $\varphi$ is exhausted by the causal contribution of some bundle of states and events (e.g., his reasons), then S does not self-determine $\varphi$

Now the disagreement between Franklin and the mainstream CTA defenders lies in P2*, particularly, in the disjunct about the proxy accounts—whether particular mental states or events can represent the agent and play the roles of the agent. Franklin argues that the proxy accounts will confront a similar dilemma as the attempt to reduce the agent into a bunch of mental states: on the first horn, if the proxy mental states are too common, then it is possible that the proxy mental states are alienated from the agent and in effect

[190] Velleman (1992, 475-476)

they cannot genuinely represent the agent; on the second horn, if the proxy mental states are too special, then it is possible that they do not always figure in the causal explanation for the agent's self-determined actions.

Franklin's reasoning is compelling. He uses an event-causal account developed by Ekstrom (1993) as an example to reveal the shortcomings of the proxy accounts. According to Ekstrom, the agent is identified with an authorized preference which is formed through a process of critical evaluation. Franklin raises two difficulties with Ekstrom's account. First, it is not clear that such authorized preferences figure in all the causal explanations for self-determined actions. Second, it is possible for an agent to act self-determinedly while against her authorized preferences. For instance, in the cases of perversity, an agent acts against her best judgment just for fun or for the sake of adventure.

Franklin's criticism applies to other proxy accounts. For example, it applies to Frankfurt's account.[191] According to Frankfurt, the agent is represented by his second-order desires; while a common objection to the account is that it is possible for an agent to be alienated from his second-order desires. This criticism also applies to Velleman's account. In Velleman's account, the agent is identified with his desire to act in accordance with reasons. The motivation for this move is that the agent cannot be alienated from the psychological items which represent him. According to Velleman, rationality is an essential part of human agency. If an agent suspends his desire to act in accordance with reasons, then the agent cannot be counted as a genuine agent. Thus, Velleman holds that it is conceptually impossible for an agent to be alienated from his desire to act in accordance with reasons.[192] Velleman's account, however, confronts the second horn of Franklin's dilemma—that the desire to act in accordance with reasons seems not always to be in the aetiology of self-determined action. If Velleman insists that the desire to act in accordance with reasons participates in every self-determined action (or full-blooded action, in Velleman's own terminology), then it will definitely overintellectualize human agency.

There is still an option for the defenders—biting the bullet by accepting the consequent of P3*. That is, denying the experience of self-determination. Franklin argues that this amounts to abandoning a significant phenomenology of self-determination. Apart from the phenomenological concern, I think there is another reason against this move. Accepting P3* means to reject the agent's having certain roles of bringing about her action—the reflective self-control and practical reasoning capacity, which is an

[191] Frankfurt 1971.

[192] Velleman (1992:479):" And in suspending the processes of practical thought, he will suspend the functions in virtue of which he qualifies as an agent. Thus, the sense in which an agent cannot disown his desire to act in accordance with reasons is that he cannot disown it while remaining an agent."

indispensable part of our conception of human agency. In this picture, the agent becomes an arena of his mental states. She no longer controls her mental states which causes her actions.

### 1.4 Producer, Coordinator and Supervisor: The Structural Account

A more plausible move to defend CTA against the Franklin's It Ain't Me Argument, I suggest, is to rebut P1*. P1* is a conditional. The antecedent is supposed to be a commitment of CTA, and the consequent is a specific view about the constituent of the self that CTA defenders have to accept if they endorse self-determination. Thus, if P1* is true, then defenders of CTA have to accept certain unpalatable consequences: either a problematic view of agent (or the self), or that self-determination is illusory.

My strategy to rebut P1* is not to argue that the conditional is false. Rather, I will argue that the antecedent of P1* is not an accurate rephrase of CTA. Therefore, defenders of CTA need not endorse the antecedent of P1*, and in effect they do not encounter the unpalatable consequences suggested by Franklin.

Then what is the problem with the antecedent of P1*? Recall the antecedent is that an agent S's causal contribution to his decision $\varphi$ is exhausted by the causal contribution of some bundle of states and events (e.g., his reasons). To put it in another way, according to this particular interpretation of CTA, the agent's roles of participating in deciding and acting has to be exclusively realized by the causing of her mental events. This idea attributes two assumptions to CTA. First, the agent's contribution to her action is exclusively causal. And relatedly, the agent's contribution has to be realized by the causing of mental states or events. Even though these two assumptions are endorsed by many defenders of CTA,[193] I will argue that they are not necessarily committed by CTA.

First of all, the contribution from an agent in action is not exclusively causal (understood in a narrow sense). Consider the case of the passive driver devised by Harry Frankfurt:

> A driver whose automobile is coasting downhill in virtue of gravitational forces alone may be entirely satisfied with its speed and direction, and so he may never intervene to adjust its movement in any way. This would not show that the movement of the automobile did not occur under his guidance. What counts is that he was prepared to

[193] E.g., Velleman (1992)

> intervene if necessary, and that he was in a position to do so more or less effectively. Similarly, the causal mechanisms which stand ready to affect the course of a bodily movement may never have occasion to do so; for no negative feedback of the sort that would trigger their compensatory activity may occur. (Frankfurt 1978, 160)

When driving downhill, the drive does not exert *actual* causal influence on the automobile. Nevertheless, he is controlling his car in the sense that he is supervising the process and he is ready to intervene with the car in case there is a need to.[194] Although this is not a very typical case of action—the agent is controlling a car rather than his body, it illuminates that the agent's roles in action are not exclusively causal. According to Frankfurt, the agent also exhibits guidance over his action, where guidance is understood as modal notion such that its effect is manifested in counterfactuals—when something goes wrong, the agent steps in and makes adjustments to her action.

Following the insights of the case, we can distinguish three roles of an agent in her action. That is, the agent as *producer, supervisor,* and *coordinator.* As a producer, the agent brings about her action. As a supervisor, the agent makes sure that the ongoing process of her action goes well. And as a coordinator, the agent makes sure that her action does not contradict her best reasons. I admit that the role of producer is wholly causal and should be realized by the causing of mental states or events in the causal-event picture. However, the role of coordinator and supervisor is not necessarily causal. What is omitted is the modal dimensions that these notions which are cashed out in counterfactuals. Particularly, if the decision or action goes well, there is no need for the agent to step in to interfere by exerting causal influence.[195]

Within the event-causal framework, there is no agent as substance to monitor her action and coordinate the various reasons regarding her action (and decision). Then what is going to play these two roles? I propose that the role of supervisor and coordinator is realized by the very structure of our psychology. I call this *the structural account*. Like Velleman and many others, I take the essential feature of human agency as being rational. But unlike Velleman, I do not think that rationality is *exclusively* achieved through practical reasoning or explicit deliberation. And relatedly, unlike Velleman and many others who

[194] Frankfurt introduces this case to rebut the standard CTA, according to which action is wholly grounded in the motivating mental states such as beliefs, desires and intentions. As mentioned in last chapter, the problem with this view is that, even though these mental states figure in the explanation for the aetiology of action, they are insufficient or sometimes irrelevant (as the case above shows) to account for the maintenance of control during the action. However, the idea behind this case is compatible with a more general CTA, that is, action can be reductively explained within an event-causal framework. In fact, Frankfurt can be friend to this more general version of CTA. For a helpful discussion of the case, see Zhu (2004).

[195] Of course, coordinator can become a causal notion. When the action or decision goes deviate, the agent need to take in and intervene.

propose the proxy accounts, I do not think that the agent requires an additional psychological item to animate practical reasoning or to propel practical reasoning to attain rationality. As a normal adult, rationality has already been built into our psychological structure. This structure includes two respects. First, our mental states have connected to one another in a largely coherent way (though not a perfectly coherent way); Second, as I have argued in Chapter 3, the agent's being reasons-responsive mainly depends on his mental states' being sensitive to reasons—that is, every single mental state on the personal level are ready to be revised in response to updated knowledge or reasons. These two respects together account for the rational structure of our psychology and play the roles of supervisor and coordinator in an action.

Here is a toy example to illustrate the structural account. Richard is feeling thirsty and has a desire to drink beer. This desire eventually transforms into his action of grasping a bottle of beer in the fridge. In bringing about the action, the desire is cooperating with a range of other mental states in Richard's mind—such as the belief that there are beers inside the fridge; that the taste of beer is good and so on. In addition, this desire is ready to be suppressed or withheld when it turns out to contradict the reasons of Richard, say, if it suddenly occurs to Richard that the doctor has said that he should not drink beer given his health condition or Richard suddenly remembers that he has an important essay to work on and should avoid alcohol (but fortunately these do not happen in the actual circumstance). Two things to be noted. First, in the whole process (from Richard's forming his desire to his action), since everything goes smoothly, no additional causal work is required except for the mental states causing the matching behaviour of Richard. Second, Richard as an agent does participate in playing the roles of supervising his action and coordinating the mental states with respect to the action. Nevertheless, Richard is not playing these roles as a substantial self over and above his mental states, nor by attributing the roles to a special proxy mental state. Rather, the roles are played by his psychological structure—that his desire to drink beer is coherently connected with other mental states and that the desire is subject to the potential updated knowledge.

To summarize, similar to the proxy accounts, the structural account does not postulate a substance-like agent to participate in her action. Different from the proxy accounts, the structural account does not entitle any proxy psychological items to *exclusively* represent the agent and to play all the roles of the agent. The structural account does not deny the existence of the special psychological items identified by those proxy accounts, such as secondary desires, desires to act in accordance with reasons, and so on. But these mental states are not more central than the normal desires in representing the agent and playing the roles of agents. In the structural account, even common mental states, such as beliefs, desires, and intentions can *partly* represent the agent in virtue of being part of the psychological structure. More importantly,

since no particular psychological items are chosen to exclusively represent the agent, we do not need to mind whether the proxy set of psychological items are too broad (in which case the agent cannot be truly represented) or too narrow (in which cases agency is over-intellectualized). In effect, the dilemma advanced by Franklin is circumvented.

### 1.5 Possible Objections to the Structural Account

There are two obvious objections to the structural account. The first one is alienation—any mental state in principle can be alienated from the agent and because of that no mental states genuinely represent the agent. And the second one is detachment—the experience that the agent *stands behind* her mental states and manipulate her mental states. I will review these two objections in order.

i) Alienation

According to the structural account, every mental state can partly represent the agent. This may fall prey to the objection of alienation—that if a mental state can be alienated from the agent, then it cannot genuinely represent the agent and play the role of the agent in her participation of action. To assess this objection, we should first clarify the notion of alienation.

In the first sense, it means that a mental state is entirely foreign to the agent—the mental state just happens to or befalls the agent without his permission. A typical case is the non-willing addict, who is suffering from an irresistible urge to take drugs. If this is what is meant by alienation, then an alienated mental state cannot represent the agent and play the roles of the agent. However, this is not a problem with my account. According to my account, a mental state (partly) represents the agent by fitting into the rational structure of the agent's psychology. An alienated mental state in this sense is neither connected coherently with other mental states of the agent, not it is sensitive to reasons. Thus, it does not play the roles of the agent.

In the second sense, an alienated mental state sometimes refers to a motive for an action which the agent later regrets, feel ashamed or guilty of performing. Velleman introduces this sense of alienation in his cases:

> Suppose that I have a long-anticipated meeting with an old friend for the purpose of resolving some minor difference; but that as we talk, his offhand comments provoke me

> to raise my voice in progressively sharper replies, until we part in anger. Later reflection leads me to realize that accumulated grievances had crystallized in my mind, during the weeks before our meeting, into a resolution to sever our friendship over the matter at hand, and that this resolution is what gave the hurtful edge to my remarks … Indeed, viewing the decision as directly motivated by my desires, and my behaviour as directly governed by the decision, is precisely what leads to the thought that as my words became more shrill, it was my resentment speaking, not I. (Velleman 1992: 464-465)

Velleman does not clearly explain why in this case the desire to severe the friendship does not genuinely represent the agent. One suggestion is that it is conceptually impossible for an agent to endorse a mental state and feel ashamed of it at the same time. And if a motive is not endorsed by the agent, it cannot represent the agent. However, unlike many proxy accounts (such as Frankfurt's), in my account, a mental state represents the agent not in virtue of being explicitly endorsed by the agent. Rather, it represents the agent by fitting into the psychological structure of the agent. Being coherently connected with other mental states is a weaker requirement than being explicitly endorsed by the agent. After all, we are not moral saints. We probably have some dark sides that we are reluctant to admit or identify as ours. Nevertheless, this does not make those motives not truly belong to us.[196]

Perhaps Velleman's point is more nuanced—that is, sometimes the motive of our action is not fully recognized by the agent—it is beyond the agent's awareness. Even if our personal-level mental states are largely coherently connected, our psychology is very often compartmentalized and sometimes there are motives or attitudes operate subliminally. The self is not a unified whole as we normally expect it to be.

Two responses in order. First, suppose that an agent does not participate when the action is driven by a subconscious motive. This causes no trouble to my account because my account only aims to articulate the agent's roles in action which are causally explained by conscious mental states.

And second, it can be argued that even a subconscious motive can represent the agent. For example, Arpaly in her *Unprincipled Virtue* has provided some interesting examples to raise the moral significance of subconscious motives. One of her examples is Huckberry Finn, a character in Mark Twin's novel, who acts against his own best judgment and helps Jim to escape slavery.[197] Finn's action is motivated by his

[196] At this point, I agree with Schlosser that unless defeating condition obtains, "actions that are nondeviantly caused by the agent's mental states and events are an expression of the agent's own agency by default". (Schlosser 2010, 26).
[197] See Arpaly 2002, 9–10.

conscience which is deeply buried in subconsciousness. Still, we find it praiseworthy because it does reflect something which truly belongs to Finn.

ii) Detachment

Franklin may point out that the structural account has difficulty with explaining the specific kind of experience related to self-determination, namely, detachment. In this experience, the self is detached from the mental states and exerting an additional causal influence on the mental states. If this experience is veridical, then there must be a substance-like agent which is detached from her mental states or mental events. However, I suggest that the experience of detachment is phenomenologically uncommon or even conceptually impossible.

First of all, I think this phenomenology of detachment is confused with a more common experience—that we *distance* specific mental states or events and reflect on them. To distance a specific mental state, an agent does not have to detach from all of her mental states. What she needs to do is just to evaluate the target mental state, using her other mental states as the frame of reference. For example, I have a desire to hang out with my friends and a competing desire to finish my essay. To distance these two desires, I am evaluating them with reference to other mental states that I already have, such as the belief that finishing the essay is more important and that I do not have enough time to finish it if I do not start working on it right now. In making the final decision, there seems to be no experience corresponding to a substance-like self over and above all the mental states. Of course, there is still an "I" who is doing the reflection. But this "I" is not detached from but constituted by various mental states which serve as the backdrop of reflection.

This relates to a conceptual point that I want to raise. How can a pure agent who is stripped of all her mental states make an adjudication or make a decision? We can think and choose because we have certain beliefs, desires, concerns and preferences. In other words, an agent can determine herself only if she has certain mental states. But what does it mean by saying 'the agent has certain mental states'? The natural interpretation is that the agent is either constituted or causally influenced by her mental states. Thus, it seems to be conceptually impossible for an agent detaching from all her mental states to make decisions on her mental states.

In summary, in this section, I have introduced some potential obstacles for reductively explaining Agent-Participation within the event-causal framework. In particular, I have critically reviewed Franklin's

*It Ain't Me* Argument which is against the reductive explanation of Agent-Participation. I have also developed an account, namely, the structural account, which shows that agent's role (as producers, supervisor and coordinators) can be fulfilled either by the causing of mental states or events or by the psychological structure of the agent.

**3. Anti-Reduction**

I have shown that how Agent-Participation obtains in an event-causal framework. Now opponents of CTA may reply that, no matter how event-causations are structured, there cannot be Agent-Participation. For events merely happen in the world, while actions by nature are active phenomena. This is the intuition of Anti-Reduction. If it is true, it will block any attempts to find agency within event-causations. This intuition is not uncommon in the literature. Here is a quote from Melden which typically represents the concern:

> It is futile to attempt to explain conduct through the causal efficacy of desire—all that can explain is further happenings, not actions performed by agents.... There is no place in this picture... even for the conduct that was to have been explained. (Melden 1961, 128–29)

Carlos Moya also provides a vivid scenario to pump our intuition of Anti-Reduction:

> … let us start with an episode that nobody would hesitate in classifying as an action, say, drinking a glass of water ...The water got into my mouth as an effect of gravity. The water getting into my mouth is a mere happening. This happening, in turn, was caused by the movement of the glass. Where is action in this? ... this movement can be said to be properly caused by my arm's and hand's movement, which in turn were caused by some muscles' contractions, which in turn were caused by some neurons' firings, and so on … Our everyday sharp distinction between actions and happenings begins to fade; it seems that we were calling 'action' what is in fact a series of causally related happenings. (Moya 1990, 3)

And more recently, Hornsby summarizes the idea in the following way:

> [Velleman (1992)] says that it is difficult to know 'how the existence and relations of .. mental states and events .., connected to one another and to external behaviour by robust causal relations, .. can amount to a person's causing something rather than merely to something's

> happening in him'. To this the answer now is simple: 'They cannot'. No compounding of states and events in the naturalistic picture from which human beings are absent could constitute someone's doing something that she intentionally does. (Hornsby 2008, 179)

These quoted passages point to the same intuition—that the picture delivered by CTA is a miracle—everything in the causal chain that passively happens to the agent will suddenly amount to the agent's doing something. Such a miracle is almost unimaginable or unthinkable.[198] Ruben once articulated the intuition as follows:

> How can activity 'emerge' from, or supervene on, the apparent passivity of events? This is, let us call it, 'the problem of passivity'. (Ruben 1997, 230)

Note that this intuition of Anti-Reduction works differently from the intuition of Agent-Participation in challenging the event-causal account of action. Agent-Participation cast doubts on the possibility of reductively explaining the agent's role in her action within an event-causal framework. By comparison, Anti-Reduction challenges the general idea of reductively explaining action by invoking the murkiness that agentive phenomena take place in a picture where all changes are fundamentally causal interactions among *non-actional* events. Thus, the intuition of Anti-Reduction serves as another source for the worry that agency disappears in the event-causal framework.

But is the picture delivered by CTA really unimaginable? Defenders from CTA can dispel the impression of unimaginability from two routes. First of all, they can highlight the active characteristics of an agent's mental life. Joseph Raz's remarks on the active/passive distinction on the mind are helpful here. Raz (1997, 223–24) proposes that many mental states (beliefs in particular) are not always something that just intrudes on us. Raz admits that mental states like beliefs are beyond the voluntary control of the agent—an agent can choose to perform a specific action while she cannot choose to believe a specific proposition. Nevertheless, he contends that many mental states are on the active side of the active/passive distinction. Raz's explanation for this claim is similar to the structural account presented above—that a well-functioning belief is usually formed as a result of being responsive to reasons.[199] An

[198] Cf., Hornsby (2004):" Besides the real phenomena that are describable using the language of alienation, there is alienation of a kind that I shall call 'unthinkable'."(174) "If you try to imagine your actions as part of the flux of events in this picture, then you will find yourself alienated from them."(174)

[199] There are still some minor differences between Raz's account and mine. Raz's discussion mainly focuses on beliefs. According to Raz, beliefs are active in virtue of being reasons-sensitivity; other mental states can be active, but in a derivative sense. For example, a desire is active partly in virtue of being informed by a reasons-sensitive belief. By comparison, in my account, a desire can be directly sensitive to reasons. In addition, for Raz, being

agent does not passively acquire a belief just as the belief somehow befalls on her. Rather, the acquisition of this belief has to be coupled with her background knowledge in the right way. And if the mental states which figure in the causal explanation of action are intrinsically active, then it becomes significantly less difficult to imagine that agentive phenomena can arise from the causation among mental states.

Another route to dispel the intuition of Anti-Reduction is by making explicit the aim of CTA. The aim of CTA is not just a reduction, but a reductive explanation. CTA is not just a claim that action amounts to a causal chain of non-actional events. Rather, CTA is a systematic program to show that certain characteristics of actions are functionally equivalent to certain causal interactions of events. Once this programme has been achieved, it is reasonable to imagine that action actually arises from event-causation. Consider a vitalist analogy. There is a vitalist about life. He makes the following claims:

> In the reductive picture of life, what we can see are just a bunch of molecules—proteins, carbohydrates, lipids, and nucleic acids. But all of these are non-living stuff. It is intuitively implausible that they can add up to life.

This argument will not convince contemporary biologists. The concept of life comes in degrees and it covers a wide range of phenomena—from simple movement in amoeba to more complicated behaviour of mammals. It is not the case that living phenomena will suddenly arise if a certain magical animate substance (say, the entelechy) adds on to the non-living molecules. There is no sharp line between living and non-living things. To reductively explain life then, then, is to specify the underlying causal mechanism of molecules realizing the functions which are essential for life—such as reproducing, metabolism, respiration, responsiveness to external stimuli and so on. Once this has been done, there seems to be nothing left about life.

The vitalist argument is analogous to the intuition of Anti-Reduction about action. One cannot simply dismiss the reductive program of action just by appealing to the intuition that non-acting events can hardly be imagined to constitute action. The real issue is whether certain actional characteristics can be reductively explained in terms of event-causal process. As suggested before, these most important actional characteristics include Control-Maintenance and Agent-Participation. Once the reductive explanation for these two characteristics has been done, there seems to be nothing left to be explained about action. If one still insists on the intuition that action arising from event-causations is unimaginable,

---

sensitive to reasons are not the only way for a belief to be active. A belief can be active through being responsive to the agent's self-deception or wishful thinking. (Raz 1997, 223–24)

then he or she is either suffering from lack of imagination, or just begging the question against the reductive program of action.[200]

Admittedly, there are two further moves to buttress the intuition of Anti-Reduction. The first is to question the tenability of reductively explaining Control-Maintenance and Agent-Participation within the event-causal framework. The second is to contend that, apart from Control-Maintenance and Agent-Participation, there is something more about action which resists reductive explanation. Since I have shown that Control-Maintenance and Agent-Participation can be reductively explained in the previous chapter and in section 1.4 of this chapter, the first move is blocked. I now review the second move.

### 3.1 Support for Anti-Reduction: Action is Essentially Subjective

Neal Judisch (2010) provides an independent explanation of why action resists reduction in terms of event-causation. According to Judisch, the obstacle is that our conception of agency essentially involves a subjective dimension. Specifically, whenever we act, we possess a special feeling of doing something, which ensures us that the action is something done by us rather than something happens to us.[201] It is this subjective aspect of action that makes it difficult to analyze action within the event-causal framework. Judisch correlates this difficulty with the problem of 'explanatory gap' in philosophy of mind, according to which, the phenomenal consciousness involves subjective characteristic, which makes it notoriously difficult to be analyzed in terms of physical properties. For Judisch, the reductivist program of action confronts a similar difficulty for action also involves a subjective aspect. He contends that the reductive program of action analyzes action under an objective perspective, which cannot do justice to the subjective aspect of action. Judisch quotes Nagel to strengthen his position:

> Something peculiar happens when we view action from an objective or external standpoint. Some of its most important features seem to vanish under the objective gaze. Actions seem no longer assignable to individual agents as sources, but become instead components of the flux of events in the world of which the agent is a part. (Nagel 1986, 110)

[200] Dennett (1991, 281–82) has made a similar vitalist analogy to deflate the idea that phenomenal properties of consciousness cannot be reduced to physical properties. Not everybody is buying his argument. However, I think this analogy will be more compelling when it runs in the realm of action. The difficulty of reducing consciousness is that phenomenal properties are difficult to be functionalized. But this difficulty has no counterpart in the reductive program of action.

[201] This echoes Carl Ginet's account of action, according which to the most important feature of basic action is the 'actish phenomenal quality'. See Ginet (1990, chap. 1).

Judisch further suggests that 'it is because the phenomenology of agency is absent from third-person conceptualizations that action theories developed from this perspective will inevitably appear to ignore something significant' and that 'what is apparently missing from the naturalistic account is a particular *phenomenal* conception of ourselves as sources of our conduct' (103). For Judisch, this ignored subjective aspect of agency can only be accommodated from 'the internal point of view' (104).[202]

Now the question is whether Judisch's explanation helps to justify the intuition of Anti-Reduction or it just explains away the intuition.[203] I suggest that Judisch's explanation, if correct, will probably make the intuition of Anti-Reduction irrelevant to rebut CTA. Firstly, according to this explanation, action resists reduction because it involves a subjective aspect. However, CTA does not aim to reductively explain the subjective experience of action. Even though some philosophers (e.g., Ginet 1990) hold that the subjective aspect is essential to action, empirical studies suggest that action can be dissociated with the experience of acting. One can act while the experience of acting goes wrong. This happens in schizophrenic patients, who can act but suffer from the delusion of alien control.[204]

Although CTA usually falls under the name of reductionism, there are actually two different reductive programs going on: the first is to reductively explain the mental (especially the phenomenal consciousness) properties into the physical properties; and the second to reductively explain actional characteristics (i.e., Control-Maintenance, Agent-Participation) in terms of event-causation. Defenders of CTA, especially those with a naturalist orientation, would admit that a more thorough understanding of human agency should include the reduction of the mental properties (the experience of agency in particular) into the physical properties. However, this task should be better left to philosophers of mind. Admittedly, for a certain version of CTA, even it is successful, it would just be one step of the entire project of naturalizing agency. Apart from cracking the hard problem of phenomenal consciousness, the naturalization of agency requires several different philosophical achievements, say, the naturalization of

[202] Judisch in his article seems to provide another explanation for the difficulty of reductively explanation—that is, the content of the phenomenology of agency speaks against an event-causal account of action (104-105). This explanation differs from his first one, according to which the subjective aspect of the phenomenology cannot be reductively explained in an event-causal framework (though Judisch in his article does not explicitly make a distinction between these two explanations). I will focus on his first explanation in this chapter. And I will deal with the challenge from the content of the phenomenology of agency in the next chapter.

[203] Judisch does not make it clear whether his explanation can help the reductive program. In some passages, he thinks the explanation helps to dispel the appearance that action cannot be reduced. For example, he borrows the phenomenal concept strategy from defenders of physicalism and claims that the 'explanatory gap' in action is just a 'conceptual gap' and should not be taken as a concern in ontology (103); in addition, he also suggests that an action theory "need not include a description of what it's like to act" (103). But in some other passages, he suggests that this explanation casts doubt on the naturalist account (104).

[204] E.g., see Spence 2001.

non-derived intentionality, and the naturalization of rationality. To fully understand agency, we definitely need input from other philosophical areas and one should not demand and expect too much from CTA.

Secondly and relatedly, according to Judisch's explanation, the problem is rooted in the distinction of external perspective and internal perspective—the agentive experience can only be done justice to from an internal, subjective, first-person perspective. However, even though the distinction of internal and external perspectives cuts across the distinction of event-causation and agent-causation, these two distinctions can be dissociated. True, within the event-causal framework, we are more inclined to view actions from an external perspective. But this is not necessary. Even within the agent-causal framework, for example, an action can be viewed either under an internal perspective or an external perspective. An action can be cashed out in terms of agent-causation from external or third-personal perspective, e.g., the agent causes her hand to rise by exercising her agential power. Or it can be cashed out in terms of event-causation from the internal or first-personal perspective, e.g., my intentions caused my hand to rise.[205] Thus, according to this explanation, even if the subjectivity of agentive experience resists any reductive analysis, we have no reason to blame the event-causal framework endorsed by CTA.

**Conclusion**

In this chapter, I have teased out two intuition sources for the problem of disappearing agency. One is the intuition of Agent-Participation, and the other is Anti-Reduction. In the first part, I have argued that Agent-Participation can be reductively explained within an event-causal account. Particularly, I have developed a positive account, namely, the structural account, to illustrate how the roles of the agent are realized by the causing of mental events or states and by the psychological structure of the agent. In the second part, I have explained away the intuition of Anti-Reduction. This amount to a partial defence of CTA against the problem of disappearing agency.

In the next chapter, I will investigate another source of the worry of disappearing agency. That is, certain kinds of phenomenology of agency seems not to fit into an event-causal framework.

[205] Perhaps some would think that the problem is not that we cannot provide an event-causation description from the first personal perspective. Rather, the problem is that whenever we try to provide such a description, it would fail to accurately capture our first personal experience of action. I think this is a fair objection and I am going to deal with it elsewhere for the limitation of space here.

## Chapter 6: Disappearing Agency and The Phenomenology of Agency

**Abstract：** In this chapter, I address the problem of disappearing agency under the interpretation according to which the causal theory of action cannot accommodate certain kinds of phenomenology of agency. I examine several kinds of phenomenology which are purported to involve content about agent-causation or non-event-causation. These include the phenomenology of acting, the phenomenology of making choice and the phenomenology of exerting effort. I argue that none of this phenomenology will pose a real challenge to the causal theory of action.

### 0. Introduction

As we have seen in chapters 4 and 5, according to the standard causal theory of action (the standard CTA henceforth), an action is a bodily movement caused by specific motivating mental states (e.g., beliefs, desires and intentions). This theory is plagued by the problem of disappearing agency: in our ordinary conception, the agent must participate in her action and play certain roles during the action; however, these roles cannot be captured by the description of action delivered by the standard CTA.

In the last chapter, I provided two interpretations of this problem. Under the first interpretation, the standard CTA fails to capture the roles of agent because the psychology of action proffered is too impoverished. According to this interpretation, there is hope to solve this problem if a more sophisticated psychological story is provided to accommodate the roles of the agent. Under the second interpretation, it is the very framework employed by CTA, namely, the event-causal framework that is insufficient to account for the roles of the agent. Under this interpretation, the problem is more severe than the first one, for it implies that the very aim of CTA, namely, reductively explaining action in terms of event-causation, is doomed to failure. Since the main aim of this thesis is to examine whether our ordinary conception of action can withstand a reduction in an event-causal framework, my focus in on the second interpretation.

In the last chapter, I have examined two motivations for thinking that the agent disappears in an event-causal description of action. There are two intuitions about action which are taken to be incompatible with the event-causal framework. The first is Agent-Participation and the second is Anti-Reduction. In defending CTA against the problem of disappearing agency, I have argued that Agent-Participation can obtain in an event-causal framework; I have also explained away the intuition of Anti-Reduction. In this

chapter, I will entertain another motivation for thinking that the agent disappears in an event-causal account—the content of our agentive phenomenology speaks against event-causation. In the first section, I provide a phenomenological reconstruction of the problem of disappearing agency. Then I will investigate several kinds of phenomenology which are allegedly speaking against the event-causal framework: section two discusses the phenomenology of acting; section three is on the phenomenology of making choices, and section four is about the phenomenology of exerting effort. I will respectively argue that all these kinds of phenomenology can fit into an event-causal account of action.

## 1. Disappearing Agency: A Phenomenological Reconstruction

### 1.1 The Roles of Agent in Action

The problem of disappearing agency is the problem that the standard CTA cannot capture the roles of an agent during her action. To assess how this problem poses a challenge to the standard CTA, we should first have a look at which roles of the agent are omitted by the standard CTA. Velleman, who originally drew the attention to this problem of disappearing agency, describes the roles of the agent in the following way:

> [I]n a full-blooded action, an intention is formed by the agent himself, not by his reasons for acting. Reasons affect his intention by influencing him to form it, but they thus affect his intention by affecting him first. And the agent then moves his limbs in execution of his intention; his intention doesn't move his limbs by itself. The agent thus has two roles to play: he forms an intention under the influence of reasons for acting and he produces behavior pursuant to that intention. (Velleman 1992, 462)

According to Velleman, the roles played by the agent in an action include mental components (making decisions or forming intentions in light of reasons) and bodily components (acting in accordance with the decisions). Following Velleman, Helen Steward also raises the problem of disappearing agency and argues that the standard CTA ignores the roles of agents. She identifies the roles as follows:

> …the agent *considers* certain reasons, *forms* an intention, *acts* on that intention. And the considering, forming, and acting that is done by the agent is therefore simply missed out of the story when we are invited to imagine that reasons by themselves causally produce intentions or that intentions by themselves causally produce bodily movements. (Steward 2012, 63, original emphasis)

Similar to Velleman, the roles specified by Steward also include mental components (considering the reasons and forming the intentions) and the bodily ones (acting). [206]

Now the question is why these roles are, supposedly, not captured by the standard CTA. One straightforward suggestion is that the standard CTA, on the one hand, says little about how those mental states preceding action are formed (e.g., the process of deliberation); on the other hand, it does not provide a sufficient account about how the agent maintains control in her action. According to the standard CTA, the agent merely initiates the action through the causing of her mental states. This does not do justice to the agent's roles in her action. For the agent not only initiates her action but also sustains and controls her action. [207]

If the problem of disappearing agency is understood in this way, then the defenders of CTA do have the resource to improve the standard CTA to capture the roles of the agent. First, they can add a story about deliberation or practical reasoning within the event-causal framework.[208] This fills the vacancy about the formation of mental states preceding action. Second, defenders can abandon the linear conception of causation adopted in the standard CTA, according to which the mental states just trigger and initiate the occurrence of bodily movement. Instead, they can equip CTA with more sophisticated causal structures, within which the mental states not only initiate but sustainedly guide the bodily movements. In Chapter 4, I have elaborated an account that can incorporate all of these dimensions of control, namely the sophisticated CTA. Specifically, in this account, the mental states and the bodily movements constitute a control system which involves feedback loops constituted with perceptual and proprioceptive feedback;[209] and even a hierarchical structure with mental states operating in a personal level and sub-personal level respectively.[210] In sum, with these adjustments, new versions of CTA have the potential to accommodate the roles of the agent in deliberating, forming intentions, and controlling the action.

---

[206] There is a difference between Velleman's position and Steward's. Velleman mainly targets on the standard CTA and thinks that certain improvements on CTA help to solve the problem. By comparison, Steward tries to apply her arguments to any possible versions of CTA. Thus, Velleman is a defender of CTA whereas Steward is an opponent of it.

[207] See Frankfurt (1978) and Searle (2003, 14) for similar concerns.

[208] Actually, this is basically how Velleman addresses the problem in his paper. Velleman thinks that the roles of agent can still be realized within the event-causal framework employed by CTA. He proposes that a psychological item should be added to the original description of action, namely, the desire to act in accordance with reasons, which can play the roles of agent. See Velleman (1992). Similar strategies to address the problem can also be found in Bratman (2001), Mele (2003) and Enc (2003)

[209] E.g., Adams and Mele (1989); Audi (1986); Thalberg (1984) Bishop (1989); Aguilar (2012). This move is crucial to handle the problem of causal deviance. See chapter Three.

[210] E.g., Pacherie (2012).

### 1.2 The Phenomenology of Agency and the Dilemma for CTA

However, the problem raised by Velleman and Steward can be interpreted in a more challenging way. The description of the roles of the agent from Velleman and Steward is compelling because it fits with our experience of acting as agents. If we look closer into the content the phenomenology of agency, we will easily find a seeming mismatch between how we *feel* about our action and how action is described in CTA. Specifically, when we are acting, we feel like it is the self rather than the mental states that issue in the action. We also have the experience that the presence of our desires and beliefs cannot fully determine the occurrence of action. Besides, when deliberating and making decisions, we have the experience of stepping back from the mental states and intervene with these mental states. These kinds of phenomenology are not easily fit into the story of action told by CTA, according to which action is fundamentally a causal process only involving mental states or mental events and bodily movements. Following this line of reasoning, the problem of disappearing agency boils down to the problem that the event-causal framework invoked by CTA cannot accommodate certain agentive phenomenological features.[211] Even if philosophers like Velleman and Steward do not intend to provide a report of phenomenology when they raise the problem, their claims still get support from the phenomenological considerations.[212] More important, this phenomenological challenge is not only a challenge specific to the standard CTA but to all versions of CTA which hinge on an event-causal framework. And I will use the notions of CTA and the event-causal account of action interchangeably in what follows.

Specifically, we can distinguish the two levels of challenge from the phenomenology. The first level of challenge can be presented as follows:

> **(First Level of Challenge)** the phenomenology of agency serves as a kind of evidence against the event-causal account of action.

For example, if we have an experience of agent-causation when we are acting or making decisions, then this kind of experience, with other things being equal, lends support to the agent-causation account of action and at the same time provides evidence against the event-causation account. Some may doubt whether the phenomenology of agency should be counted as evidence for or against a specific account of action. After all, we should care about what action *actually is* but not what action is *felt to be*. Unless we

[211] Note that this is not the only interpretation for the problem of disappearing agency. In last chapter, I have identified several possible interpretations for this problem.

[212] Agent-causation libertarians are often motivated by the phenomenological concerns (e.g., O'Connor 1995). Given that the problem of disappearing agency is often taken to support the agent-causation account of action., it strikes me that so few philosophers have explicitly associated the problem of disappearing agency with the topic of phenomenology of agency. Notable exceptions are Judisch (2010), Pereboom (2015) and Franklin (2018)

have reasons to believe that our subjective experience of action actually reveals the ontological structure of action, we do not need to consider phenomenology for a true account of action.

Then there is a weaker while more tenable challenge engendered by the phenomenology. As Horgan et al. (2003) suggest, the phenomenology of agency sets up certain veridicality conditions which can only be fulfilled by a specific theory of action. Thus, if it turns out that there are veridicality conditions that cannot be fulfilled by the event-causation account, then it may turn out that either the event-causation account is false or that our experience is spurious. Here comes the second level challenge:

> **(The Second of Level Challenge)** the phenomenology of agency sets up certain veridicality conditions which are not satisfied by CTA.

For example, if our phenomenology of action involves content about agent-causation, it follows that the phenomenology represents reality only if our action is agent-caused. In other words, the phenomenology is illusory if the action is event-caused.

Confronting the second level of challenge, if proponents of CTA are going to bite the bullet by conceding that the identified phenomenology of agency is not accommodated by CTA while insisting that the phenomenology of agency does not reveal the real causal structure of action, then the proponents of CTA have to admit that at least a significant part our phenomenology of agency is illusory in the sense that the veridicality conditions of phenomenology are never satisfied. In what follows, I will talk about the challenge of phenomenology only in this second level interpretation.

Still, some philosophers may think that the second level challenge would not cause much trouble to CTA. After all, we have already accepted that our perceptual experience, in general, is fallible and does not always reflect reality. These philosophers would suggest that we can live with the illusory experience without being uneasy. For example, Carl Ginet, who defends a non-causal account of action, seems to accept with no hesitation that the phenomenology of acting (presented as agent-caused) is by and large illusory.[213] And Helen Beebee in an unpublished manuscript about the phenomenology of free will argues that it would *not* be especially puzzling or incoherent for someone who has the libertarian experience and endorses compatibilism. She provides the reasons as below:

> We're all used to things not being as they appear. Sticks look bent in water; redness really does look like it's an intrinsic quality of objects; tables really don't seem to be mostly empty space; and you can feel hungry and then, just as you're reaching for a Tim

[213] See Ginet 1990, 13.

> Tam, realize that you can't possibly be hungry because you only just had lunch. (Beebee, unpublished manuscript)

Beebee compares the illusion of phenomenology of agency to the fallibility of our general perceptual experience. Though Beebee's concern is in the phenomenology of free will, her argument of analogy applies to general phenomenology of agency as well.

However, there are important disanalogies between the phenomenology of agency and the fallibility of general perceptual experience. The general perceptual experience, even being fallible, only breaks down in limited circumstances or fail to represent the reality in some of the aspects; we normally believe that our general perceptual experience is reliable in normal situations (otherwise general scepticism will arise). For example, we all know that the visual experience sometimes can be false (e.g., Müller-Lyer illusion) but we still trust it most of the time. The reliability of the general perceptual experience can explain why we can live with the falsehood without being incoherent. The phenomenology of agency, however, if shown to be illusory, will be illusory on an extremely large scale. It means every time we exercise our agency, there is an accompanying phenomenology which does not represent reality. Besides, the phenomenology of agency arguably plays an important role in shaping the conceptions of ourselves—as a free and responsible agent. This conception serves as the foundation of our moral practices. If it turns out that our agentive experience is not veridical, it may lead to the consequence that a large part of our moral practices require revision. Thus, even if the illusory experience of agency cannot be a knock-down argument against CTA, it would nevertheless make CTA a dramatically less attractive or even counter-intuitive position to accept. Also, the defenders of CTA should bear the burden to explain why we would have this wholesale illusory experience.

In the remainder of this chapter, I will explore the phenomenology of agency in detail and argue that it is compatible with the event-causation account of action. My primary aim is to investigate whether our experience of agency sets up veridicality conditions which cannot be satisfied by an event-causal account of action. To achieve this aim, I will first identify and spell out several kinds of agentive experience which would be potentially regarded as speaking against the event-causation account. I will in particular discuss three different kinds of phenomenology which may cause trouble for the event-causation account, namely, the phenomenology of acting, the phenomenology of making choices and the phenomenology of exerting effort. I will argue that the veridicality conditions set up by these kinds of phenomenology can all be met by an event-causation account of action.

## 2. The Phenomenology of Acting

Some philosophers suggest that the very phenomenology of acting may lend support to the idea of agent-causation. That is, when we are acting, we seem to experience that our actions are directly caused by ourselves, rather than are caused by our mental states such as beliefs and desires. For example, Horgan et al. (2003) write that "[y]our phenomenology presents your own behavior to you as having yourself as its source, rather than (say) presenting your own behavior to you as having your own occurrent mental events as its source". In a more recent article, Horgan writes that "[y]ou experience your arm, hand, and fingers as being moved *by you yourself*—rather than experiencing their motion either as fortuitously moving just as you want them to move, or passively experiencing them as being caused by your own mental states... You experience your behavior as *caused* by yourself, rather than experiencing it as caused by *states* of yourself." (Horgan 2011, 79, original emphasis) These reports of our experience seem to suggest that our phenomenology of acting involve a specific veridicality condition—action is not caused by mental events, but caused by the self. This veridicality condition can hardly be satisfied by an event-causal account of action.[214] These verdicts can be construed as the second level challenge to CTA: even if defenders of CTA doubt the truth-conduciveness of this experience, it also makes them face a dilemma—if the phenomenology represents reality, then CTA is false; if not, then our phenomenology of acting is illusory (and so does not meet relevant veridicality conditions). To fully establish this reasoning, we have to specify this kind of phenomenology and reconstruct the arguments in detail. Specifically, phenomenological features captured by Horgan and colleagues can be used to rebut CTA in two ways—one directly and the other indirectly. In what follows, I will reconstruct two corresponding versions of argument based on the phenomenology of acting.

### 2.1 The Direct Argument

The phenomenology of acting can be deployed in a direct way to rebut CTA: one gets the conviction that her action is caused by the self (as an irreducible substance) rather than caused by the mental events or mental states directly through the phenomenology of acting. I call this *the direct argument.* According to this argument, the phenomenology of action directly presents to us the content about agent-causation. If

[214] Even though the phenomenology highlighted by Horgan and colleagues can be used to support an agent-causal account of action, it should be note that in their article they do not take sides in the metaphysical nature of agency. Horgan and colleagues entertain different proposals to interpret this phenomenology and they are open to the option that the phenomenology is compatible with event-causation or neutral to event-causal condition or agent-causal condition. Horgan in his later writing even suggests that attending to phenomenology is insufficient to set up metaphysical conditions which are relevant to the debate in free will and philosophy of action. See note 218 below.

this argument is sound, the defenders of CTA face a dilemma—either our phenomenology of acting reflects reality and CTA is false, or our phenomenology of acting is illusory. The key premise of the direct argument is that the phenomenology of acting directly delivers to us content about agent-causation (or at least content about non-event-causation). I think there are strong reasons to doubt this premise. To reveal the problem with this premise, I should first talk about how people usually discern a concept.

There are two basic ways to discern a concept—we can discern it phenomenologically or we can discern it conceptually. To discern a concept phenomenologically, there must be certain distinct qualitative content attached to the particular instance of that concept. This is how we typically do when we are distinguishing pain from itch, sweet from sour, red from blue and so on—these notions are all associated with a distinctive quality which can be delivered to us either through introspection or perception. Alternatively, we can discern a concept in a conceptual way—typically by providing definitions of the concept or outlines the relevant characteristics of that concept with language. We can distinguish a prime number from a composite number for they involve different mathematical definitions. Of course, these two ways of discerning concepts can cut across one another. For example, we can distinguish a circle from a square either in a phenomenological way or in a conceptual way. Now the question is whether we can discern the concepts of agent-causation/event-causation in a phenomenological way. This question boils down to another one—whether there are distinctive phenomenal qualities pertaining to the concepts of agent-causation/event-causation. I suggest the answer is probably negative. To clarify, I am not suggesting that there is no phenomenal quality associated with the concept of causation in general. Rather, I am suggesting that there is no fine-grained phenomenal quality that helps to distinguish agent-causation from event-causation. [215]

First, even though a distinction can be made between agent-causation and event-causation, the distinction is usually made in a conceptual way. Specifically, for event-causation, the relata of the causal interactions must exclusively be events. For agent-causation, the cause in a causal relation must be the agent understood as an irreducible substance. Accordingly, agent-causation is a special case of substance-causation. Since the distinction between event and substance is crucial to the distinction of event-causation and agent-causation, it is worth entertaining the possibility that event and substance can be distinguished phenomenologically. But these two notions are usually distinguished in a conceptual way. Philosophers will say the following about the difference between substance and event:

[215] David Hume famously proposes that causal relation can never be reflected in our perceptual experience. Unlike Hume, I am open to the question whether causation can be experienced or not. What I am defended here is much less radical than Hume's claim—even if there is a phenomenology of causation about action, we cannot tell it is agent-causation or event-causation merely from the phenomenology.

> In ordinary language, we have different predicates for substance and event: we say that a substance exists, and an event occurs. Besides, substance and event bear different relations to space and time. Substances locate at specific spatial points—they remain relatively stable and identifiable within a certain period. By comparison, events are temporal particulars are more often taken as changes rather than persistence…[216]

This characterization of the difference between event and substance seems to be too abstract to be reflected in our experience. Perhaps one can propose that the experience about changing and persisting is sufficient to distinguish event and substance in a phenomenological way since substance is usually associated with persistence and event with change. This proposal does not work. For we also categorize an unchanging datable particular as an event (e.g., it is a description of event that 'the apple remains fresh during the cool afternoon').

There is a collateral argument for the lack of distinctive phenomenal quality pertaining to event-causation and agent-causation. In our perceptual experience, we can use substance-causation statements or event-causation statements interchangeably to report the same causal phenomena. For instance, we can say that "the rushing of the ball causes the window's breaking", which is an event-causation claim; or we can equally say that "the ball breaks the window", which is a substance-causation claim. However, both claims are made based on the same observation.[217] This suggests that our observational experience is neutral to the choice between event-causation and substance causation. Likewise, we can reasonably expect that our introspective experience does not distinguish event-causation and agent-causation.

In summary, I have shown that the directive argument hinges on the premise that there is distinctive phenomenal quality pertaining to agent-causation and event-causation. I have argued that this premise is spurious.[218] Therefore, it is untenable to hold that our phenomenology of acting can directly present us with content involving agent-causation condition.

---

[216] The points are extracted from Casati and Varzi (2020).You can find more detailed discussion and a complete survey of the literature in their article.

[217] This argument is inspired by Beebee's unpublished manuscript.

[218] Horgan (2007; 2011) has also provided sophisticated arguments for the claim that the we cannot ascertain the metaphysical conditions of the phenomenology of acting just by introspection (the conditions concerned by Horgan are broader—including agent-causation, causal closure and determinism). Horgan's basic idea is the following: forming judgments of the metaphysical conditions of the phenomenology require conceptual competence; and people's conceptual capacity is limited. Thus, people cannot immediately exercise the conceptual capacity to form the judgment while introspecting to the phenomenology. Horgan's argument is based on the acceptance of the idea that there is distinctive phenomenal quality pertaining to agent-causation, an idea which I have argued against. In my argumentation, no matter how powerful one's conceptual capacity is, she cannot extract agent/event-causation condition from her agentive experience. However, I am happy to adopt Horgan's argument as a backup in case someone insists on the existence of distinctive quality about agent-causation.

## 2.2 The Indirect Argument

But now another question arises: if there is no content about agent-causation directly delivered to us, then why do so many people, including philosophers, come to get the seeming impression that the phenomenology of acting involves content about agent-causation? I propose an answer as follows. Even if we cannot *directly* recognize any presentational content pertaining to agent-causation in our phenomenology of acting, at least we do find an experiential difference between the cases of action and the uncontroversial cases in which our body is definitely event-caused to move by mental states and events (e.g., a pain causes me to flinch). Specifically, we do have salient experience when our bodily movements are definitely event-caused by our mental states. For example, one experiences that a state of fear can cause her body to start; or that the intense pain in one's body causes her to scratch aimlessly. Call these the *uncontroversial cases of mental event-causation*. But our phenomenology of acting is significantly different from what we feel in the uncontroversial cases of mental event-causation: in particular, when acting, we have a special *sense of control*; we have a feeling that we are in charge of our own actions; this feeling cannot be found in the uncontroversial cases of event-causation.[219]

I propose that this is how we get the ostensible impression that the phenomenology of acting involves a condition of agent-causation. This impression is based on comparison and inference. First, we find a significant difference between the phenomenology of action and the phenomenology of the uncontroversial cases of mental-event causation (typically, the former cases involve a sense of control). And second, from this phenomenal difference, we further infer that the phenomenology of acting involves content about agent-causation.[220] Besides, this inference can serve as another argument to establish the condition of agent-causation from the phenomenology of acting**:** even though we cannot directly recognize phenomenal content about agent-causation by attending to the phenomenology of acting, we can indirectly infer that the phenomenology involves agent-causation by comparing it to the

[219] I am not suggesting that the sense of control can exhaust the difference between the two kinds of phenomenology. There may be others—such as a sense of voluntariness. See Horgan et al. 2003.

[220] Horgan et al. (2003) also introduce the phenomenology of acting in a negative/contrastive way—they do not directly cash out what it is like in action; rather they cash out what it is *not* like in action. Later they propose two possible ways in describing the phenomenology of acting—*experiencing non-event causation* and *not experiencing event-causation*. Horgan (2007) further argues that it is the latter that should be the appropriate description of our phenomenology of acting. However, I think their distinction should be made with more qualifications. It is not the case that we experience non-event causation in action. Nor it is not the case that we do not experience event-causation *simpliciter* in action. Rather, it is the case that we find a phenomenal difference between action and uncontroversial cases of mental event-causation. And I suggest this phenomenal difference is the sense of control. Plausibly, the sense of control is neutral to the characterization of event-causation or agent-causation.

uncontroversial cases of mental event-causation. I call it the indirect argument, which can be reconstructed as an inference to the best explanation as follows:

> (P1) There is a phenomenal difference between the cases of action and the non-controversial cases of mental event-causation.
>
> (P2) The best explanation for this phenomenal difference is that the phenomenology of acting involves content about agent-causation

Therefore,

> (C) the phenomenology of acting involves a veridicality condition of agent-causation.

If this argument is sound, then defenders of CTA will again confront a dilemma—either our phenomenology of acting is illusory in terms of failing to meet the veridicality conditions, or CTA is false. Since P1 is convincing, the better way to rebut this argument is to attack P2.

In section 2.1, I have argued that there is probably no distinctive phenomenal quality pertaining to agent-causation or event-causation. That conclusion can apply here to reject P2 as well: if there is no distinctive phenomenal quality pertaining to agent-causation, then we cannot directly recognize the agent-causation content in our phenomenology, nor can we indirectly infer there is any. Thus, the phenomenal difference between action and the uncontroversial cases of mental event-causation should not be explained by the existence of agent-causation content. But someone can just turn the table around and state the following words: since there is a phenomenal difference between action and the uncontroversial cases of mental event-causation, there must be distinctive phenomenal quality pertaining to agent-causation and event-causation. To block this move, I will provide other explanations available for the phenomenal difference.

First, in the uncontroversial cases of mental event-causation, usually, the causes of the resultant behaviour are either bodily sensations (e.g., pain, itch) or affective states (e.g., fear, anger). These mental states uniformly involve salient specific phenomenology. By comparison, the motivating states which causally explain the actions, such as beliefs, desires and intentions, usually lack *salient specific* phenomenal character. This partly explains the phenomenal difference between the two types of cases. There are two qualifications for my claim.

By specific phenomenology of the mental state, I mean the phenomenal character which is specific to the attitudinal component rather than the content of the mental state in question. Sometimes the content of

the mental states may involve phenomenal character.[221] Suppose I believe that tomorrow is a sunny day. This belief may involve phenomenal character in virtue of its content, such as the mental images of sunshine, blue sky. But all these are not the specific phenomenology in question. For those mental images are sensory experience contingently attached to the content of this belief: I can have the same mental image without the attitudinal component of the belief, say I just entertain that thought; or I can just simply have the belief without those mental images.[222]

To claim the mental states such as beliefs do not involve any specific phenomenal character will be controversial and risky. Actually, the question of whether intentional mental states (e.g., believes and desires) involve specific phenomenology triggers a grand debate under the theme of cognitive phenomenology. However, what I require is a weaker claim—that these motivating mental states lack *salient* specific phenomenal character. Even if the motivating mental states involve some specific phenomenal character, it must be elusive and impoverished. This claim is justified by the status of the on-going debate in cognitive phenomenology.[223] If there were salient specific phenomenal character attached to beliefs and desires, it would not trigger so much controversy among philosophers.

My second explanation focuses on the sense of control, a significant phenomenal aspect which distinguishes actions from the uncontroversial cases of mental event-causation is the sense of control. There is a burgeoning literature on the sense of control from several disciplines (e.g., philosophy, psychology and motor cognition). In the literature, a popular model to account for this sense of control is the comparator-predictor model. This model is based on a widely shared theoretical framework in motor cognition. According to this framework, the agent obtains control in her action through a computational-representational process. Roughly, the control system on the one hand sustainedly makes predictions about the output of bodily motion; on the other, it continuously receives feedback information about the output, compares it with the predictions and regulate the body motions accordingly. The basic hypothesis of the comparator-predictor model is that the sense of control arises if, in comparison, the predicted output matches the actual output. Empirical research supports this idea.[224] Even if this model may not be the final story of how the sense of control comes about,[225] the take-home message is that the sense of control may probably represent certain properties of the bodily motion related to the maintenance of

[221] Or sometimes not. Just think about those abstract beliefs such as the belief that 2 is a prime number.
[222] I think similar conclusions apply to the motivated kinds of desires. Nagle (Nagel 1970, 29) has introduced a useful distinction between motivated desires and nonmotivated desires—motivated desires are arrived at after deliberations, while unmotivated desires just assail us and they are like appetite and emotions (e.g., hunger). For example, the desire to get a PhD in philosophy is probably a motivated desire, and hunger is an unmotivated one. Following this distinction, I suggest that the motivated desires are usually lack salient specific phenomenology.
[223] For a review of the debates in cognitive phenomenology, see Bayne and Montague (2011)
[224] For a summary of the empirical evidence for the comparator-predictor model, see Bayne & Pacherie (2007).
[225] For a recent view of the shortcomings of this model, see Mylopoulos & Shepherd (forthcoming)

control (e.g., the matching of the perceptual feedback and the intended motion). As I have shown in section 1.1 of this chapter, and more extensively in Chapter 4, the maintenance of control in action can be realized with an event-causal structure involving feedback loops, thus the sense of control involved in the phenomenology of acting may probably be neutral to the event-causation and agent-causation.

To wrap things up, first, I have argued that the notions of agent-causation and event-causation, as abstract metaphysical notions, are unlikely to involve distinct phenomenal quality. Therefore, we cannot directly recognize phenomenal content specific to agent-causation or event-causation, nor can we indirectly infer any content specific to agent-causation or event-causation. Apart from that, I have provided two explanations for the phenomenal difference between actions and the uncontroversial cases of mental event-causation. The first explanation is that the motivating mental states which causally explain the action do not involve salient specific phenomenal character. The second explanation is that there is a sense of control involved in actions; according to recent studies, this sense of control probably represents the properties related to the maintenance of control in bodily motions. I conclude that the phenomenology of acting does not set up a veridicality condition of agent-causation.

### 3. The Phenomenology of Making Choices and the Sense of Indetermination

The phenomenology of making choices is another source for invoking agent-causation. Particularly, when we have the experience of making choices, we will easily come across a sense of indetermination. As Holton writes:

> It is quite compatible with a given set of beliefs and desires either that we choose one way or that we choose another. That, of course, is part of what makes choice an action: we are not pushed along by our beliefs and desires. (Holton 2006, 4)

In other words, the choice made by the agent is felt not to be determined by any mental states prior to the decision. Here is the reasoning. This sense of indetermination may be incompatible with an event-causation account of action: since the prior mental states do not guarantee the occurrence the final decision, [226] there must be a gap between the mental states and the final decision or the resultant action. The opponents of the event-causation account may hold that this gap can only be filled by agent-causation. Therefore, the event-causation account is false or that this phenomenology of making choices is illusory.

[226] I will shortly explain what 'guarantee' means below.

Is the argument convincing? I think not. The essential claim of the argument is that the sense of indetermination involved in choice-making is incompatible with the event-causation account. To evaluate this argument, it is important to specify this sense of indetermination. Although Holton in the quoted text has characterized the conception of indetermination in clear language, that "we are not pushed along by our beliefs and desires", it seems to me that something more can be said. In particular, the quote seems to involve two slightly different senses of indetermination: one is epistemological— the mental states such as beliefs and desires do not constitute sufficient or compelling reasons for the agent to act; another is psychological— the mental states such as desires and beliefs do not necessitate the chosen action. In what follows, I will give a close look at these two senses of indetermination and argue that both are compatible with the event-causation account.

### 3.1 The Epistemological Sense of Indetermination

Recall that Holton states that '[i]t is quite compatible with a given set of beliefs and desires we choose one way or we choose another.' This can mean that the mental states of the agent such as beliefs and desires do not constitute compelling reasons for a specific option such that the agent cannot decide whether to act or not. This echoes Holton's idea that the requirement of choice comes only "when the questions what to do arise".[227] That is to say, choice-making experience only arises in specific contexts in which there is something uncertain concerning the next move.[228] I refer to this uncertainty as the epistemological sense of indetermination.

Here is an account of how this epistemological sense of indetermination comes about. An agent's mental states at a specific moment may favour several different courses of action. On the one hand, this can happen because of competing motivations. For example, I may want to write my essay and play my Nintendo Switch at the same time. And both acts can be rationalized by my current mental states from my subjective perspective. On the other hand, sometimes our mental states favour different options because the given situations involve more than sufficient resources. Simply put, we have too many choices available to fulfil our desires. For example, I am now feeling hungry and would like to go to a restaurant for a meal. Near my living place, there is a KFC and a McDonald's. Since I do not have particular

[227] Holton (2006, 3). Accordingly, he also argues that people do not have to choose in every action. In everyday life, most of human actions and activities take place due to habits or automatic process. And thanks to this fact, we can spare ourselves a lot of cognitive resources.
[228] C.f. Mele (2003, chap. 9) argues that decision making is to resolve uncertainty.

preferences for either restaurant, my mental states can be both compatible with the option of KFC or McDonald's.

Now the crucial question is whether the epistemological sense of indetermination involves content of non-event-causation or agent-causation? I think the answer is probably not. For this sense of indetermination mainly concerns an epistemological relation rather than causal relation. In what follows I will develop the arguments in more detail. Specifically, I will distinguish two cases of epistemological indetermination. The first one concerns the uncertainty that can be resolved in deliberation about the choice-making process. And the second concerns the uncertainty that cannot be resolved even after deliberation. Particularly, it is the second case that might invite an agent-causation interpretation.

As proposed by Holton, choice-making only happens in the contexts where the question of what to do next arises. Arguably, this question arises when the agent's current mental states do not provide sufficient reasons to form the intention of the next act. In such a situation, there is an epistemic gap between the agent's current mental states and her potential concomitant courses of action. To fill this gap, the agent then needs to deliberate. A typical deliberation for choice may involve several steps:

1) different possible courses of action and the anticipated consequences are entertained;

2) the relevant reasons for the possible courses of action are found;

3) the options are ranked and the judgment about the best option is made;

4) an intention to act is formed.

Arguably, all these steps can be either viewed as mental actions or part of mental actions. In addition, all these steps in principle can be cashed out in event-causation terms. In some situations, the uncertainty confronting the agent can be resolved after the process of thinking (if the uncertainty is due to a theoretical problem) or deliberation (if it is due to a practical one). For example, Mike is taking a math exam. There is a one-option question which requires Mike to choose the correct answer:

"Which combination of three side lengths can make a triangle? **A**. 1 2 3, **B**. 2 2 4, **C** 2 3 5, **D**. 3 4 5"

Apparently, this question has a right answer to be reached. But Mike is not good at mathematics. When Mike first reads the question, he cannot figure out the answer immediately. There is an epistemological gap between his mental states and his decision on the answer. Mike then performs the process of thinking—retrieving the relevant mathematical knowledge from his memory, considering each option, doing some mental calculations. Fortunately, the uncertainty is overcome by his thinking. Notably,

throughout the process—from the arising of uncertainty to the resolution of it, there is no need to appeal to agent-causation.[229]

The more challenging cases are those involving uncertainty that cannot be resolved even after a deliberation. As Holton (2006; 2010) suggests, when an agent is making a choice, she does not have to reach a judgment about what is the best option to choose. In these cases of choosing, the agent has competing options but cannot decide the best one. This happens typically when the agent does not have a superior reason for any specific options. For example, Mike is considering how to spend his summer break—he can either do an internship or attend a summer school. It is a hard choice and Mike does not find a superior reason to favour either option because both seem to be attractive and make sense to him. Going to a summer school may facilitate his academic performance while having an internship may prepare him better for the job market. The reasons favouring either option counterbalance each other. Now consider another case in which Mike is thinking about buying a red cap or a brown cap in a clothing store. He is not making a difficult choice in the sense that there is no significant consequence involved. However, he still cannot see a superior reason for either option because he does not have any colour preferences. In both cases, Mike may have to pick a choice without a judgment about what is the better/best option to choose. Otherwise, he would confront the predicament of the Buridan's ass and cannot make further moves.

In the above two cases, the mental states do not provide sufficient or compelling reasons for the agent to act even after deliberation. To reuse the metaphor, there is a gap between the mental states and the final decision, a gap that cannot be bridged by the agent's deliberation. It seems that this gap has to be filled by something beyond the mental states or the mental events. This is why it is tempting to assume that it is the agent rather than the mental states which settle the final decision to act. If this reasoning is correct, then the phenomenology of epistemological indetermination suggests something which may conflict with the event-causation account and even lend support to the agent-causation account.

I have two replies in order. First, this reasoning does not reach a conclusion about phenomenology. In this reasoning, the condition of agent-causation is taken as a postulate to account for the outcome of choice-making. However, this reasoning does not show that agent-causation is *presented* in our phenomenology. In the process of making choices without a judgment, we do not experience any apparent mental states which serve as decisive reasons favouring a final decision about what the best option is; nor

[229] Although this is a case of theoretical thinking, not practical deliberating or reasoning. However, a parallel case about practical reasoning can be easily imagined.

do we experience any apparent agent-causal process that leads to the final decision.[230] Actually, the process is more like a black box to us—we do not have many clues about how a decision is made by introspecting to the experience.

Second, there is a more plausible story about how a choice can be made without judgment. When the agent is making a choice, many cognitive mechanisms are operating beyond her conscious level. There is a huge literature in psychology and cognitive sciences indicating that agent's choices and decisions may be influenced by factors beyond awareness. In cognitive psychology, there is a research paradigm known as the choice-blindness, which has got fruitful results indicating that people do not have many clues about what is going on in their mind (especially the reasons or the motives for their decisions) when they are making decisions. For example, in an experiment conducted by Johansson and his colleagues, subjects are shown a pair of female faces and are asked to choose the one they find more attractive.[231] The subjects are sometimes asked to provide reasons for their decisions. Among some of the trials, the experimenter deliberately changes the image chosen by the subjects and then present the unchosen one to the subjects and ask them about the reasons. It turned out that few of the subjects can recognize this mismatch.

Return to the problem of epistemological indetermination. An agent can adopt different methods to choose without judgment. For example, he can flip a coin or use some similar tools to generate a random outcome to facilitate the decision. More often the agent resolves the uncertainty just by picking an option *randomly*, which means that when the choice is made, the agent is not consciously aware of any superior reasons. There may be indeterminism or even agent-causation involved in this process. But this is not necessarily so. As mentioned, there may be some mechanism in the sub-personal level that can account for the outcome of choice-making.[232] For example, in the case of choosing caps, Mike eventually picks up the brown one because deep in his sub-consciousness he prefers brown. Or in the case of choosing how to spend the summer break, Mike eventually chooses to apply for the internship because he was primed by an article which he read a few days ago on Facebook about how an internship is crucially helpful for a young job seeker while Mike himself is not aware of the influence of this article. The lesson is that the uncertainty between the mental states of which we are aware and the final decision can be bridged by subconscious mental states. Since the phenomenology of epistemic indetermination does not distinguish the agent-causation explanation and this sub-consciousness explanation, we cannot conclude that this phenomenology sets up the veridicality condition of agent-causation.

---

[230] And as argued in 2.1, an agent-causal process is unlikely to be experienced. But the main argument here can run independently of the argument in 2.1.
[231] Johansson et al. (2005)
[232] Very roughly, if a mental state is in the personal level, the agent can verbally report the content of it; if a mental state is in the sub-personal level, the agent cannot verbally report the content of it.

### 3.2 The Psychological Sense of Indetermination

The epistemological sense of indetermination discussed above has not exhausted the experience of indetermination in the choice-making process. There is another sense of indetermination, to which I refer as the psychological sense of indetermination. It means that in a choice-making process, our current mental states are not felt to necessitate any of our decisions or actions. To illustrate the idea, we can think of the opposite where the behaviour is felt to be determined by the mental states: a drug addict's act to take more drugs seems to be forced by his irresistible urge. In normal cases, no matter how intense a desire is, it would not drive the agent's choice or action as an irresistible urge does.

The psychological sense of indetermination is different from the epistemological sense of indetermination discussed above. There can be the psychological sense of indetermination without the epistemological one: even if my current mental states do provide sufficient and compelling reasons for a specific option, phenomenologically, it is possible that I do not choose that option. Just think about the cases of *akrasia* in which the agent fails to act in accordance with her best judgment because of the weakness of will; this can also happen in the cases of perversity, e.g., a teenager acts against her best judgement because she wants to be outlandish.[233]

This psychological sense of indetermination may also be taken to be incompatible with the event-causation account of action. The reasoning goes like this: in our experience, the motivating mental states do not determine the decision or the occurrence of action. Therefore, something else (probably, the agents as non-reducible substance) must be invoked to explain the final decision or the occurrence of action.

However, this reasoning is problematic for it conflates causation with necessitation. Even though the motivating mental states are not experienced as pushing or forcing the action, it is false to infer that the mental states do not cause the action. The ordinary conception of causation is more relaxed. Specifically, when we say that A causes B, we do not mean that whenever A happens, B necessarily happens. Rather, what we mean is that with certain normal conditions held fixed, when A happens, B would happen. This conception of causation does leave space for akrasia or perversity: the agent's mental states can cause her decision and action if nothing abnormal happens, or the mental states may fail to issue in the action—say, if some whimsical motivations befall the agent and make her suddenly change her mind. And this story is compatible with the event-causal framework.

[233] James Lenman once told me that there is a British idiom to describe this phenomenon, which is 'bloody-mindedness'.

In conclusion, in the phenomenology of making choices, neither the epistemological sense of indetermination nor the psychological sense of indetermination involves a veridicality condition of agent-causation.

### 3.3 Activeness and Detachment

Apart from the sense of indetermination, the phenomenology of making choices involves another component that may speak against an event-causal account. That is, a sense of active engagement. Some defenders of CTA might leave us the impression that choice-making is just a process of different desires competing against each other and the strongest desire wins. For example, Davidson once wrote, "if reasons are causes, it is natural to suppose that the strongest reasons are the strongest causes."[234] However, our own experience of choice-making tells a different story: usually, the agent actively engages in the whole process of making choices—considering the rationality of the beliefs and desires, and resist the irrational temptation if necessary. We are not like to be mechanically caused by our mental states to reach a decision. Consider the quotes from O'Connor:

> [Agent-causation] theory is appealing because it captures the way we experience our own activity. It does not seem to me (at least ordinarily) that I am caused to act by the reasons which favour doing so; it seems to be the case, rather, that I produce my own decisions in view of those reasons…(O'Connor 1995, 196)

However, I do not think this is a real challenge for the CTA. This phenomenological feature only indicates that choice-making is something that an agent does. Thus, to accommodate this phenomenological feature, a defender of CTA needs to provide a more sophisticated account of choice-making, according to which choice-making can be regarded as a kind of mental action which is something the agent does. It is reasonable to expect that this account can be given within an event-causal framework.[235]

Another relevant phenomenological feature of choice-making is what I call the detachment. Specifically, when we are making choices or deliberating, we have the feeling such that our selves are detached from our mental states and can consider, weigh, or even manipulate the relevant mental states.

[234] Davidson (1980/2001, xvi).
[235] For example, see Mele (2003, chap. 9)for an actional account of practical deciding which is compatible with an event-causal account. See also Shepherd (2015)for an account about how agentive control is obtained in the process of practical deliberation.

This picture of the detached self is very popular in the literature. It is endorsed by opponents and even defenders of CTA. Here are some examples:

> When an agent reflects on the motives vying to govern his behaviour, he occupies a position of critical detachment from those motives; and when he takes sides with some of those motives, he bolsters them with a force additional to, and hence other than, their own. His role must therefore be played by something other than the motives on which he reflects and with which he takes sides. (Velleman 1992, 476-477)

> The image of the agent directing and governing is, in the first instance, an image of the agent herself standing back from her attitudes, and doing the directing and governing. (Bratman 2005, 33–34)

> It is this seeming experience of myself [when making hard choices] as playing a causal role over and above the causal role of my desires and beliefs that suggests I exercised the power of self-determination. (Franklin 2018, 182)

> When you deliberate, it is as if there were something over and above all of your desires, something which is *you*, and which *chooses* which desire to act on. (Korsgaard 1996, 100, original emphasis)

In this picture, the deliberating agent not only substantially differs from her mental states, but also has the power to manipulate and adjudicate her mental states. If the picture of detachment is an accurate description of our phenomenology of deliberation and making choices, it will follow that the phenomenology involves presentational content which is not satisfied by the event-causal account of action. According to this picture, the deliberational process cannot be reduced to the causal interactions among mental states; there is the self, which is "over and above" the mental states, participating in deliberation and choice-making. Accordingly, by endorsing this picture, defenders of CTA may either abandon the event-causal framework or concede that the phenomenology of making choices is illusory at least in some respects.

I now raise some worries with this picture. First, it is not clear that this is a reliable report of our experience of making choices or that it is just another philosophical myth. Suppose I am making a hard choice—deciding whether I should go to graduate school for a PhD in philosophy or go to law school for a JD. Am I just deliberating like a purely rational homunculus[236] in the sense that I am free of any

[236] Actually, the paradox of homunculus, which is supposed to attack the computationalism in philosophy of mind, can apply here as well: according to this picture, in a deliberation process, the self behind his mental states is manipulating, choosing and deciding on his mental states. However, in choosing and deciding his mental states, the

attitudes and that I manipulate my beliefs and desires just like doing mathematical calculations? I think this is probably not the case or at least not the usual case. A more realistic picture is like this: I seem to be tortured by the separated selves—sometimes I feel more inclined to the philosophy PhD when the imagination happens to me in which I am reading interesting philosophical texts or discussing philosophy; suddenly, it occurs to me that some early-career philosophers have complained on a website about how competitive the job market is and what mental issues they are suffering because of that. Then I feel more inclined to the option of law school. However, the thought starts to strike me that I am going to spend days and nights reading tedious legal cases. In a while, Kant's words about the heavenly stars above and the moral laws within fill in my mind and again make my goosebumps pop up all over my body…[237]The point I am going to make from the case is that, during deliberation, the self cannot come apart from the mind. One's thinking self is *always* intertwined with his mental states— his beliefs, desires, attitudes and emotions.

However, some may wonder, if this picture of detachment is rare if not impossible in our phenomenology, then where it comes from; and why, if it does not reflect our experience of making choices, it is so widely shared among philosophers. I have two suggestions for this concern.

First, the idea of detachment probably comes from a metaphor in our ordinary language: in our daily talk, we do sometimes say that an agent can stand behind her mental states, reflect on her mental states, and exert causal influence over and above her mental states. This kind of language leaves us the ostensible impression that there is an agent which is stripped of her mental states. However, if we look closely at it, this picture appears to be difficult to make sense of. Just think about the question how can a pure agent who is stripped of all her mental states make an adjudication or make a decision? We can think and choose because we have certain beliefs, desires, concerns and preferences. In other words, an agent can determine herself only if she has certain mental states. But what does it mean by saying 'the agent has certain mental states'? The most natural interpretation is that the agent is either constituted or causally influenced by her mental states. Both interpretations are compatible with an event-causal framework, as already argued above.

Second, this idea may be rooted in our ability to reflect on our certain mental states. Admittedly, we can *distance* ourselves from specific mental states. One, for instance, can reflect on a specific desire to eat a chocolate bar given that he is trying to lose weight. And one can reconsider his preconceptions of

---

self must have already got some kinds of mental states (e.g., preferences, and beliefs); otherwise, choosing is a non-starter.

[237] See also Arpaly (2002, chap. 1) where she provides some vivid example to argues that 'cool hour' deliberation is not the normal case.

homosexuality if he has a chance to interact with homosexual people, etc. However, we should not conflate this ability with the detachment. Distancing oneself from a specific mental state does not mean that there is a self that is free of all his mental states. We can re-evaluate a specific mental state just because we have already had a bunch of mental states in operation. For example, one can abandon his desire to eat the chocolate bar because he already holds the belief that the chocolate bar contains more calories than he needs. Or one can abandon his bias towards homosexual because he is appreciating the value of fairness. To summarize, the ability to distance oneself from certain mental states does not require that there is a detached agent who is stripped of all mental states.

**4. The Phenomenology of Exerting Effort**

The final kind of phenomenology to investigate in this chapter is the phenomenology of exerting effort. Brent (2012) argues that CTA has difficulties with accommodating the phenomena of effort in action. The notion of effort brings trouble to the event-causation account from two respects. First, if the action becomes too tough to continue such that agent needs to exert effort to continue, the previous motivational mental states seem no longer to be sufficient to causally explain and rationalize the effortful action in question. Brent uses the example of running a marathon to illustrate this point. The agent has some previous motivational mental states such as beliefs, desires and intentions to finish the marathon. However, when the running becomes too harsh, the previous motivational mental states are no longer sufficient to rationalize the continuance of the running and something different has to be invoked to explain the sustaining of the action. He suggests what can explain the sustaining of the action is the agent's exerting effort in exercising the relevant bodily capacity (a capacity to perform basic bodily actions such as moving one's limbs). This carves out a distinctive role for the agent, namely, exerting effort.

This worry can be dispelled if we allow new mental states to arise during the action. Even if the initial mental states may not be enough to sustain the task in question, the agent can acquire new mental states that are responsible for the completion of the action being performed. Reconsider the case of running a marathon. When the task becomes too harsh, the agent may have a motivation to give up. If, however, the agent has a strong commitment to finishing the marathon, this commitment would give rise to new intentions to continue the running and overcome the difficulties. And the interactions of these mental states can in principle be characterized purely in terms of event-causation.

A more difficult question is from the phenomenology of exerting effort. When we are exerting effort, we feel like that the provider or the source of effort seems to exclusively be the agent but not mental states

such as beliefs, desires and intentions. As Brent writes, "such effort is something that the agent does or exerts, and it is thus something that we attribute directly to them, and so it cannot be reduced to or identified with the motivational factors that figure in the standard story of action."[238] In a word, the phenomenology of exerting effort seems to involve content about agent-causation.

To assess this challenge, we should have a more detailed specification of the phenomenology of exerting effort. Though the experience is common, providing a precise description of it is difficult. Different authors have articulated this experience in different ways. Bayne and Levy, for example, characterizes it as "the experience of needing to invest energy and will-power in our actions" and he suggests that we normally experience this feeling in situations such that "[t]he world's resistance to our actions, coupled with our limited success in changing it".[239] And Horgan et al. write that the phenomenology of exerting effort involves an "element of uncertainty about success".[240] Besides, Pacherie (2007) has distinguished the experience of mental effort and the experience of physical effort. In a word, there seems not to be a uniform characterization of the phenomenology of exerting effort.

I hold that there are two important remarks about effort. First, effort comes in degrees. There is no sharp distinction between effortful actions and effortless actions. If I can run 1km at a certain speed without (feeling of exerting) effort, then perhaps I need some effort to run 1.5km at the same speed, and even more to run 2km and so on.[241] Or we can just say that every action requires effort to some extent, even though in many situations the effort is too trivial to be experienced. Second, effort comes in different shapes: I require effort to do 50 push-ups; I require effort to crack a difficult mathematical problem; I require effort to clear all stages of a video game; I require effort to eat up a bowl of spicy ramen (for I am not quite into spicy food). All of these effortful activities must involve different experiences of exerting effort.

Based on these two remarks, I propose that there is no *sui generis* phenomenology corresponding to exerting effort; rather, the phenomenology of exerting effort is a *complex*—there is a common base with various toppings. The base of the phenomenology of exerting effort is the phenomenology of acting which is introduced in section 2. This is because exerting effort and performing an action is not conceptually separable. Whenever we are exerting effort to do something, we are doing something with effort. Besides, the phenomenology of exerting effort can have different toppings, depending on the task being performed. For example, if I am exerting effort in running, I may experience certain kind of

[238] Brent (2012, 25)
[239] Bayne and Levy (2006, 57).
[240] Horgan et al. (2003: n7).
[241] See Bayne and Levy (2006, 58) for a similar remark.

difficulties related to running, such as fatigue and being short of breath. If I am exerting effort in doing push-ups, I may experience the contracting of my muscle; if I am exerting effort in cracking mathematical problems in an exam, I may experience focus and my working memory being overloaded; if the consequence of the task is important, then I may even experience anxiety or nervousness.

If, as proposed, the phenomenology of exerting effort is a complex—with a common base and various toppings, then it is probably the base, namely the phenomenology of acting, that are mainly responsible for the ostensible impression that the phenomenology of exerting effort involves content about agent-causation. In particular, when we are exerting effort, we are usually acting in the situation of counteracting difficulties. This makes the phenomenology of acting even more salient than usual. Accordingly, we can more easily get the ostensible impression of agent-causation from the phenomenology. However, as I have argued in Section 2, the phenomenology of acting does not involve presentational content pertaining to agent-causation. Therefore, the phenomenology of exerting effort does not involve presentational content pertaining to agent-causation either.

**Conclusion**

In this chapter, I have reconstrued the problem of disappearing agency as a phenomenological challenge to CTA. Specifically, there are certain kinds of phenomenology of agency seeming to involve veridicality conditions that cannot be satisfied by an event-causal account of action. I have reviewed three kinds of phenomenology, namely, the phenomenology of acting, the phenomenology of making choices, and the phenomenology of exerting effort. These seem to be the strongest cases for setting up the veridicality condition of agent-causation. I have shown that all these kinds of phenomenology can be compatible with an event-causal account of action. This chapter and the previous chapter conclude my defence of CTA against the problem of disappearing agency. That is, the event-causal framework does not make the agent disappear.

## Conclusion

I started the thesis by presenting the two ways in which we look at human agency: the manifest and scientific images. Within the manifest image, agency is understood through agents' motivations, choices and reasons, while in the scientific image it is understood through causes, laws of nature and mechanisms. I have defended an account of free agency, which, on the one hand, is able to meet our ordinary expectations delivered by the manifest image of human agency; on the other, is compatible with plausible assumptions form the scientific image—determinism and universal event-causation. This account maintains that:

> (i.) Free Agency (which is required by moral responsibility) is the ability to respond to reasons.
>
> (ii.) To exercise free agency is to act for reasons; while acting for reasons can be captured within an event-causal framework.

To develop this account, I have defended a reasons-responsiveness theory and a causal theory of action (CTA) respectively. I have also shown that these two theories can be integrated—both theories can be cashed out in terms of causal properties and modal properties within the event-causal framework. This implies that the important features of human agency, from basic purposive actions to free and responsible actions can be explained in terms of counterfactual conditionals and event-causation. Besides, I have shown that these two theories can fully be combined: CTA serves as the basis for the reasons-responsiveness theory, and the reasons-responsiveness theory helps CTA to circumvent the problem of disappearing agency. More importantly, these two theories together provide a promising approach to reconciling the two images of human agency mentioned in Chapter 1. With the integrated account, we can preserve most of our conceptions from the manifest image of human agency even if we accept the metaphysical ideas (determinism and universal event-causation) exacted from the scientific image.

Below is a summary of the key argumentative moves and the achievements in the thesis.

In the first part of the thesis, I have defended a reasons-responsiveness view that an agent's responsible agency is grounded in her ability to respond to reasons. This view includes two basic ideas, first, reasons-responsiveness consists of the ability to do otherwise, conditionally understood (defended in Chapter 2); second, reasons-responsiveness, as a modal property, can figure in the causal explanation for the agent's free and responsible action (defended in Chapter 3). This account can accommodate the direct intuition from the Frankfurt-Style cases that an agent's ability to do otherwise is not necessary for moral responsibility. It can also accommodate a picture of agency inspired by the Frankfurt-Style cases, that

acting freely and responsibly is reflected in the causal history of the actions. Thus, my account can handle both the Challenge of Unnecessity (namely, reasons-responsiveness, consisting of the ability to do otherwise, is not necessary for moral responsibility) and the Challenge of Irrelevance (namely, reasons-responsiveness, as an unmanifested modal property, is irrelevant to responsible action) from the Frankfurt-Style cases.

There are several additional achievements in developing the account of reasons-responsiveness. First, the account helps to resolve the current schism between the leeway compatibilism and source compatibilism. This division is caused by a seeming tension of two ideas: on the one hand, freedom is a modal property which is to be cashed out in counterfactuals; on the other hand, inspired by the Frankfurt-Style cases, freedom is wholly grounded in the actual-causal history. According to my account, these two ideas are not conflicting but are about different aspects of freedom (Chapter 2).

Second, this account motivates and outlines a new argumentative strategy against incompatibilism. That is, to urge the incompatibilists for a story about how incompatibilist free will is explanatorily relevant to the agent's action (Chapter 3).

Third, this account provides a story about how reasons-responsiveness can figure in the causal explanation for the action. This paves the wave to incorporate the causal theory of action into the reasons-responsiveness theory (Chapter 3).

The second part of the thesis is devoted to defending the causal theory of action. The causal theory of action confronts two main difficulties—the problem of causal deviance and the problem of disappearing agency.

In Chapter 4, I have provided a strategy to deflate the problem of causal deviance. Rather than seeking sufficient and necessary conditions for non-deviance, I have proposed that the crucial move to solve the problem is reductively explaining the agent's maintaining control within an event-causal framework. I have provided a functional analysis of control, to which I referred as the gradualist account of control. Based on this account, I have shown that the agent's maintaining control can be realized within an event-causal framework (relying on the sophisticated CTA developed by Peacocke and Bishop). And more importantly, according to this account, control is a notion that comes in degrees and it is multi-faceted. Since no sharp distinction between control and non-control can be drawn, there is no need to provide sufficient and necessary conditions for a non-deviant causal chain in solving the problem.

In this thesis, I have diagnosed that the worry of disappearing agency comes from different sources. First, it is due to two different intuitions about action, namely, Agent-Participation (that the agent has to

participate in her action) and Anti-Reduction (that agency as active phenomena cannot merely arise from event-causation). Second, it is due to several kinds of phenomenology of agency.

In Chapter 5, I have shown that Agent-Participation can be reductively explained within the event-causal framework. Specifically, I have developed the structural account (with recourse to the previous discussion about reasons-responsiveness) showing that the agent's roles in action can be realized within the event-causal framework. I have also explained away the intuition of Anti-Reduction.

In Chapter 6, I have provided a phenomenological reconstruction of the problem of disappearing agency. I have investigated several kinds of phenomenology of agency which purportedly speak against the event-causal account. I have argued that none of them involves content about agent-causation or non-event-causation. Thus, Chapter 5 and 6 constitute a complete defence of CTA against the problem of disappearing agency.

The integrated theory advanced in this thesis can be applied to address further issues in the future. For example, there is a burgeoning literature suggesting the importance of proprioception and perceptions in action. The event-causal account of action developed in the thesis has the potential to incorporate the perceptual and proprioceptive states into the causal structure of action and provide a more complete picture of agentive control. In addition, there are challenges to reasons-responsiveness theories from the empirical researches in psychology. These researches suggest that people's choices and behaviour are influenced by situational factors which are beyond conscious awareness. The reasons-responsiveness account developed in this thesis can help to illustrate the nature of this challenge and evaluate the extent of this challenge.

**Bibliography**


Aguilar, Jesús H. 2012. “Basic Causal Deviance, Action Repertoires, and Reliability.” *Philosophical Issues* 22 (1): 1–19.

Armstrong, D. M. 1983. *What Is a Law of Nature?* Cambridge University Press.

Arpaly, Nomy. 2002. *Unprincipled Virtue: An Inquiry Into Moral Agency*. *Unprincipled Virtue*. Oxford University Press.

Ayer, A. J. 1954. “Freedom and Necessity.” In *Philosophical Essays*, 3–20. New York: St. Martin’s Press.

Bayne, Timothy J., and Neil Levy. 2006. “The Feeling of Doing: Deconstructing the Phenomenology of Agency.” In *Disorders of Volition*, edited by Natalie Sebanz and Wolfgang Prinz. Cambridge: MIT Press.

Bayne, Timothy J., and Elisabeth Pacherie. 2007. “Narrators and Comparators: The Architecture of Agentive Self-Awareness.” *Synthese* 159 (3): 475–91.

Bayne, Timothy J., and Michelle Montague. 2011. “Cognitive Phenomenology: An Introduction.” In *Cognitive Phenomenology*, edited by Timothy J. Bayne and Michelle Montague. Oxford University Press. https://www-oxfordscholarship-com.sheffield.idm.oclc.org/view/10.1093/acprof:oso/9780199579938.001.0001/acprof-9780199579938-chapter-1.

Beebee, Helen. 2004. “Causing and Nothingness.” In *Causation and Counterfactuals*, edited by L. A. Paul, E. J. Hall, and J. Collins, 291–308. Cambridge, MA, USA: MIT Press.

———. n.d. “Is the Phenomenology of Free Will Relevant to Its Metaphysics?” *Unpublished Manuscript*.

Beebee, Helen, and Alfred Mele. 2002. “Humean Compatibilism.” *Mind* 111 (442): 201–23.

Bishop, John. 1989. *Natural Agency: An Essay on the Causal Theory of Action*. Cambridge University Press.

Björnsson, Gunnar, and Karl Persson. 2012. “The Explanatory Component of Moral Responsibility.” *Noûs* 46 (2): 326–54.

Brand, Myles. 1984. “Intending and Acting: Toward a Naturalized Action Theory.” *Journal of Philosophy* 84 (1): 49–54.

———. 1989. “Proximate Causation of Action.” *Philosophical Perspectives* 3: 423–42.

Bratman, Michael. 1987. *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

———. 2001. “Two Problems About Human Agency.” *Proceedings of the Aristotelian Society* 101 (3): 309–326.

———. 2005. “Planning Agency, Autonomous Agency.” In *Personal Agency*, edited by Stacey James, 35–61. Cambridge University Press.

Brent, Michael. 2012. “Effort and the Standard Story of Action.” *Philosophical Writings* 40: 19–27.

Brink, David O., and Dana K. Nelkin. 2013. “Fairness and the Architecture of Responsibility.” In David Shoomaker, ed. *Oxford Studies in Agency and Responsibility Volume 1*, edited by David Shoemaker:*284-313*. Oxford University Press.

Casati, Roberto, and Achille Varzi. 2020. “Events.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2020/entries/events/.

Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. OUP USA.

Chisholm, Roderick M. 1964. “Human Freedom and the Self.” The Langley Lecture, 1964 (University of Kansas). Reprinted in J. Feinberg and R. Shafer-Landau, eds., *Reason and Responsibility: Readings in Some Basic Problems of Philosophy*, 11th Edition (Wadsworth, 2002), 492- 99

———. 1966. “Freedom and Action.” In *Freedom and Determinism*, edited by Keith Lehrer. Random House.

———. 1978. “Comments and Replies.” *Philosophia* 7 (3–4): 597–636.

Clarke, Randolph. 2009. “Dispositions, Abilities to Act, and Free Will: The New Dispositionalism.” *Mind* 118 (470): 323–351.
———. 2017. “Free Will, Agent Causation, and ‘Disappearing Agents.’” *Noûs*, 76–96.
———. 2019. “Agent Causation and the Phenomenology of Agency.” *Pacific Philosophical Quarterly* 100 (3): 747–64.
Cleland, Carol, and Christopher Chyba. 2010. “Does ‘Life’ Have a Definition?” In *The Nature of Life: Classical and Contemporary Perspectives from Philosophy and Science*, edited by Carol E. Cleland and Mark A. Bedau, 326–39. Cambridge: Cambridge University Press.
Davidson, Donald. 1963. “Actions, Reasons, and Causes.” *The Journal of Philosophy* 60 (23): 685–700.
———. 1970. “Mental Events.” In *Experience and Theory*, edited by L. Foster and J. W. Swanson, University of Massachusetts Press. Reprinted in Davidson (2001), 207-225
———. 1973. “Freedom to Act.” In *Essays on Freedom of Action*, edited by Ted Honderich. Routledge. Reprinted in Davidson (2001), 63-81
———. 1974. “Psychology as Philosophy.” In *Philosophy of Psychology*, edited by S. Brown. Harper & Row. Reprinted in Davidson (2001), 229-239
———. 1978. “Intending.” *Philosophy of History and Action* 11: 41–60. Reprinted in Davidson (2001).
———. 2001. *Essays on Actions and Events*. Oxford: New York.
———. 2004. *Problems of Rationality*. Oxford University Press.
Dennett, Daniel. 1991. *Consciousness Explained*. Penguin Books.
———. 2004. *Freedom Evolves*. Penguin Books.
Ekstrom, Laura W. 1993. “A Coherence Theory of Autonomy.” *Philosophy and Phenomenological Research* 53 (3): 599–616.
———. 2000. *Free Will: A Philosophical Study*. Westview.
Enc, Berent. 2003. *How We Act: Causes, Reasons, and Intentions*. Oxford University Press.
Fara, Michael. 2008. “Masked Abilities and Compatibilism.” *Mind* 117 (468): 843–65.
Fischer, John Martin. 1994. *The Metaphysics of Free Will: A Study of Control*. Blackwell.
———. 2007 “Compatibilism.” In Fischer, Kane, Pereboom and Vargas. 2007. *Four Views on Free Will (Great Debates in Philosophy): 5*. Illustrated edition. Malden, MA; Oxford: John Wiley & Sons.
———. 2012. *Deep Control: Essays on Free Will and Value*. Oxford University Press USA.
———. 2015. “Responsibility and the Actual Sequence.” In *Oxford Studies in Agency and Responsibility: Volume 3*, edited by David Shoemaker. Oxford University Press. https://oxford-universitypressscholarship-com.sheffield.idm.oclc.org/view/10.1093/acprof:oso/9780198744832.001.0001/acprof-9780198744832-chapter-7.
———. 2018. “The Freedom Required for Moral Responsibility.” In *Virtue, Happiness, Knowledge: Themes from the Work of Gail Fine and Terence Irwin*, 216–33.
Fischer, John Martin, and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.
Ford, Anton. 2011. “Action and Generality.” In *Essays on Anscombe’s Intention*, edited by Anton Ford, Jennifer Hornsby, and Frederick Stoutland. Harvard University Press.
Frankfurt, Harry G. 1969. “Alternate Possibilities and Moral Responsibility.” *The Journal of Philosophy* 66 (23): 829–39.
———. 1971. “Freedom of the Will and the Concept of a Person.” *The Journal of Philosophy* 68 (1): 5–20.
———. 1978. “The Problem of Action.” *American Philosophical Quarterly* 15 (2): 157–62.
Franklin, Christopher Evan. 2015. “Everyone Thinks That an Ability to Do Otherwise Is Necessary for Free Will and Moral Responsibility.” *Philosophical Studies* 172 (8): 2091–2107.
———. 2016. “If Anyone Should Be an Agent-Causalist, Then Everyone Should Be an Agent-Causalist.” *Mind* 125 (500): 1101–31.
———. 2018. *A Minimal Libertarianism: Free Will and the Promise of Reduction*. Oxford University Press.

Ginet, Carl. 1990. *On Action*. Cambridge England; New York: Cambridge University Press.
Goetz, Stewart C. 1988. "A Noncausal Theory of Agency." *Philosophy and Phenomenological Research* 49 (2): 303–16.
Goldman, Alvin I. 1970. *A Theory of Human Action*. Princeton University Press.
Haddock, Adrian. 2005. "At One with Our Actions, but at Two with Our Bodies: Hornsby's Account of Action." *Philosophical Explorations* 8 (2): 157–172.
Hoefer, Carl. 2016. "Causal Determinism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2016/entries/determinism-causal/.
Holton, Richard. 2006. "The Act of Choice." *Philosophers' Imprint* 6: 1–15.
———. 2010. "Disentangling the Will." In *Free Will and Consciousness: How Might They Work?* edited by Al Mele, Kathleen Vohs, and Roy Baumeister, 82. Oxford University Press.
Honderich, Ted. 1988. *A Theory of Determinism*. Oxford University Press.
Horgan, Terry E., John L. Tienson, and George Graham. 2003. "The Phenomenology of First-Person Agency." In *Physicalism and Mental Causation*, edited by Sven Walter and Heinz-Dieter Heckmann, 323. Imprint Academic.
Horgan, Terry. 2007. "Agentive Phenomenal Intentionality and the Limits of Introspection." *PSYCHE: An Interdisciplinary Journal of Research On Consciousness* 13.
———. 2011. "The Phenomenology of Agency and Freedom: Lessons From Introspection and Lessons From Its Limits." *Humana. Mente* 15: 77–97.
———. 2015. "Injecting the Phenomenology of Agency into the Free Will Debate." In *Oxford Studies in Agency and Responsibility: Volume 3*, edited by David Shoemaker. Oxford University Press. https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198744832.001.0001/acprof-9780198744832-chapter-3.
Hornsby, Jennifer. 1980. *Actions*. Routledge & Kegan Paul.
———. 2004. "Agency and Actions." *Royal Institute of Philosophy Supplement* 55: 1–23.
———. 2008. "Agency and Alienation." In *Naturalism In Question*, edited by M. de Caro and D. MacArthur, 173–87. Cambridge, USA: Harvard University Press.
Hume, David. 1748. *An Enquiry Concerning Human Understanding*. Reprinted in 2008. Edited by Peter Millican. New edition. Oxford University Press.
Hyman, John. 2015. *Action, Knowledge, and Will*. Oxford University Press.
Inwagen, Peter van. 1983. *An Essay on Free Will*. Oxford University Press.
Irwin, Terence H. 1980. "Reason and Responsibility in Aristotle." In *Essays on Aristotle's Ethics*, edited by Amélie Oksenberg Rorty, 117–155. University of California Press.
Jackson, Frank, and Philip Pettit. 1990. "Program Explanation: A General Perspective." *Analysis* 50 (2): 107–17.
Jeannerod, Marc. 1997. *The Cognitive Neuroscience of Action*. The Cognitive Neuroscience of Action. Malden: Blackwell Publishing.
Johansson, Petter, Lars Hall, Sverker Sikström, and Andreas Olsson. 2005. "Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task." *Science* 310 (5745): 116–19.
Judisch, Neal. 2010. "Bringing Things About." In *New Waves in Metaphysics*, edited by Allan Hazlett, 91–110. New Waves in Philosophy. London: Palgrave Macmillan UK.
Kane, Robert. 1996. *The Significance of Free Will*. Oxford University Press USA.
———. 2005. *A Contemporary Introduction to Free Will*. Oxford University Press.
Keil, Geert. 2007. "What Do Deviant Causal Chains Deviate From?" In *Intention, Deliberation and Autonomy*, edited by Christoph Lumer and Sandro Nannini, 69–90. Ashgate.
Kim, Jaegwon. 1998. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. MIT Press.
Korsgaard, Christine M. 1996. *The Sources of Normativity*. Cambridge University Press.
Kratzer, Angelika. 1977. "What 'Must' and 'Can' Must and Can Mean." *Linguistics and Philosophy* 1 (3): 337–55.

Lehrer, Keith. 1968. “Cans without Ifs.” *Analysis* 29 (1): 29–32.
Levine, Joseph. 1983. “Materialism and Qualia: The Explanatory Gap.” *Pacific Philosophical Quarterly* 64 (4): 354–61.
Levy, Yair. 2013. “Intentional Action First.” *Australasian Journal of Philosophy* 91 (4): 705–18.
Lewis, David. 1976. “The Paradoxes of Time Travel.” *American Philosophical Quarterly* 13 (2): 145–52.
———. 1986. “Causal Explanation.” In *Philosophical Papers Vol. II*, edited by David Lewis, 214–240. Oxford University Press.
Lowe, E. J. 2008. *Personal Agency: The Metaphysics of Mind and Action*. Oxford University Press.
Maier, John. 2018. “Abilities.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2018/entries/abilities/.
Mayr, Erasmus. 2011. *Understanding Human Agency*. Oxford University Press.
McKenna, Michael. 2013. “Reasons-Responsiveness, Agents, and Mechanisms” In *Oxford Studies in Agency and Responsibility: Volume 3*, edited by David Shoemaker. Oxford University Press. https://www-oxfordscholarship-com.sheffield.idm.oclc.org/view/10.1093/acprof:oso/9780199694853.001.0001/acprof-9780199694853-chapter-7.
McKenna, Michael, and David Widerker, eds. 2003. *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities: Essays on the Importance of Alternative Possibilities*. Routledge.
Mckitrick, Jennifer. 2005. “Are Dispositions Causally Relevant?” *Synthese* 144 (3): 357–71.
Melden, Abraham I. 1961. *Free Action*. Routledge.
Mele, Alfred R. 1992. *Springs of Action: Understanding Intentional Behavior*. Oxford University Press.
———. 2003. *Motivation and Agency*. Oxford University Press.
Moore, Michael S. 2010. “Renewed Questions about the Causal Theory of Action.” In *Causing Human Actions: New Perspectives on the Causal Theory of Action*, edited by Jesús H. Aguilar and Andrei A. Buckareff, 27–43. The MIT Press.
Moya, Carlos. 1990. *The Philosophy of Action: An Introduction*. Polity Press.
Mylopoulos, Myrto, and Joshua Shepherd. forthcoming. “Agentive Phenomenology.” In *Oxford Handbook of the Philosophy of Consciousness*, edited by Uriah Kriegel. Oxford University Press.
Nagel, Thomas. 1970. *The Possibility of Altruism*. Princeton University Press.
———. 1986. *The View From Nowhere*. Oxford University Press.
Nelkin, Dana K. 2005. “Freedom, Responsibility and the Challenge of Situationism.” *Midwest Studies in Philosophy* 29 (1): 181–206.
———. 2011. *Making Sense of Freedom and Responsibility*. Oxford University Press.
O’Brien, Lilian. 2012. “Deviance and Causalism.” *Pacific Philosophical Quarterly* 93 (2): 175–196.
O’Brien, Lucy. 2017. “Actions as Prime.” *Royal Institute of Philosophy Supplements* 80 (July): 265–85.
O’Connor, Timothy. 1995. “Agent-Causation.” In *Agents, Causes, and Events: Essays on Indeterminism and Free Will*, edited by Timothy O’Connor. Oxford University Press.
———. 2000. *Persons and Causes: The Metaphysics of Free Will*. Oxford University Press USA.
Pacherie, Elisabeth. 2007. “The Sense of Control and the Sense of Agency.” *PSYCHE: An Interdisciplinary Journal of Research On Consciousness* 13: 1–30.
———. 2012. “Action.” In *The Cambridge Handbook of Cognitive Science*, edited by Keith Frankish and William Ramsey, 92–111. Cambridge University Press.
Peacocke, Christopher. 1979. *Holistic Explanation: Action, Space, Interpretation*. Oxford: Clarendon Press.
Pereboom, Derk. 2001. *Living Without Free Will*. Cambridge University Press.
———. 2014. *Free Will, Agency, and Meaning in Life*. Oxford University Press.
———. 2015. “The Phenomenology of Agency and Deterministic Agent Causation.” In *Horizons of Authenticity in Phenomenology, Existentialism, and Moral Psychology: Essays in Honor of*

*Charles Guignon*, edited by Hans Pedersen and Megan Altman, 277–94. Contributions To Phenomenology. Dordrecht: Springer Netherlands.
———. 2018. "On Carolina Sartorio's Causation and Free Will." *Philosophical Studies* 175 (6): 1535–1543.
Pynchon, Thomas. 1973. *Gravity's Rainbow*. Viking Press.
Raz, Joseph. 1997. "The Active and the Passive: Joseph Raz." *Aristotelian Society Supplementary Volume* 71 (1): 211–228.
Ruben, David-Hillel. 1997. "The Active and the Passive: David-Hillel Ruben." *Aristotelian Society Supplementary Volume* 71 (1): 229–46.
Ryle, Gilbert. 1949. *The Concept of Mind*. London: Hutchinson. Reprinted in 1990, Penguin Books.
Sartorio, Carolina. 2015. *Sensitivity to Reasons and Actual Sequences*. In *Oxford Studies in Agency and Responsibility: Volume 3*. edited by David Shoemaker. Oxford University Press. https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198744832.001.0001/acprof-9780198744832-chapter-6.
———. 2016. *Causation and Free Will*. Oxford University Press UK.
———. 2018a. "Replies to Critics." *Philosophical Studies* 175 (6): 1545–1556.
———. 2018b. "Replies to Critics." *Teorema: Revista Internacional de Filosofía* 37 (1): 107–22.
Schlosser, Markus E. 2007. "Basic Deviance Reconsidered." *Analysis* 67 (3): 186–94.
———. 2010. "Agency, Ownership, and the Standard Theory." In *New Waves in Philosophy of Action*, edited by A. Buckareff, J. Aguilar, and K. Frankish, 13–31. Palgrave-Macmillan.
———. 2013. "Conscious Will, Reason-Responsiveness, and Moral Responsibility." *Journal of Ethics* 17 (3): 205–232.
Schwitzgebel, Eric. 2019. "Belief." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/belief/.
Searle, John R. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.
———. 2003. *Rationality in Action*. Cambridge, Mass.: MIT Press.
Sehon, Scott R. 1997. "Deviant Causal Chains and the Irreducibility of Teleological Explanation." *Pacific Philosophical Quarterly* 78 (2): 195–213.
Sellars, Wilfrid. 1962. "Philosophy and the scientific image of man", In *Science, Perception, and Reality*. edited by Robert Colodny Humanities Press/Ridgeview. 35-78. Reprinted in 2007. *In the Space of Reasons: Selected Essays of Wilfrid Sellars*. edited by Kevin Scharp and Robert B. Brandom (eds). 369-408. Harvard University Press.
Shepherd, Joshua. 2014. "The Contours of Control." *Philosophical Studies* 170 (3): 395–411.
———. 2015. "Deciding as Intentional Action: Control Over Decisions." *Australasian Journal of Philosophy* 93 (2): 335–351.
Smith, Michael. 2003. "Rational Capacities, Or: How to Distinguish Recklessness, Weakness, and Compulsion." In *Weakness of Will and Practical Irrationality*, edited by Sarah Stroud and Christine Tappolet, 17–38. Oxford: Clarendon Press.
———. 2009. "The Explanatory Role of Being Rational." In *Reasons for Action*, edited by David Sobel and Steven Wall, 58–80. Cambridge: Cambridge University Press.
Spence, Sean A. 2001. "Alien Control: From Phenomenology to Cognitive Neurobiology." *Philosophy, Psychiatry, and Psychology* 8 (2–3): 163–172.
Steward, Helen. 2012. *A Metaphysics for Freedom*. Oxford University Press.
Strawson, Galen. 1994. "The Impossibility of Moral Responsibility." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 75 (1/2): 5–24.
Strawson, Peter F. 1962. "Freedom and Resentment." *Proceedings of the British Academy* 48: 1–25.
Tännsjö, Torbjörn. 2009. "On Deviant Causal Chains – No Need for a General Criterion." *Analysis* 69 (3): 469–73.
Taylor, Richard. 1966. *Action and Purpose*. New York: Humanities Press.

Thalberg, Irving. 1984. “Do Our Intentions Cause Our Intentional Actions?” *American Philosophical Quarterly* 21 (3): 249–60.
Vargas, Manuel. 2013. *Building Better Beings: A Theory of Moral Responsibility*. Oxford University Press.
Velleman, J. David. 1992. “What Happens When Someone Acts?” *Mind* 101 (403): 461–481.
Vetter, Barbara, and Romy Jaster. 2017. “Dispositional Accounts of Abilities.” *Philosophy Compass* 12 (8): e12432.
Vihvelin, Kadri. 2000. “Freedom, Foreknowledge, and the Principle of Alternate Possibilities.” *Canadian Journal of Philosophy* 30 (1): 1–23.
———. 2004. “Free Will Demystified: A Dispositional Account.” *Philosophical Topics* 32 (1/2): 427–450.
———. 2013. *Causes, Laws, and Free Will: Why Determinism Doesn’t Matter*. *Causes, Laws, and Free Will*. Oxford University Press.
Wallace, R. Jay. 1996. *Responsibility and the Moral Sentiments*. 185. Harvard University Press.
———. 1997. “Review of The Metaphysics of Free Will: An Essay on Control.” *The Journal of Philosophy* 94 (3): 156–59.
Watson, Gary. 1975. “Free Agency.” *The Journal of Philosophy* 72 (8): 205–20.
———. 2012. “Responsibility and the Limits of Evil: Variations on a Strawsonian Theme.” In *Free Will and Reactive Attitudes: Perspectives on P.F. Strawson’s “Freedom and Resentment,”* edited by Michael McKenna and Paul Russell. Ashgate Publishing, Ltd.
Whittle, Ann. 2010. “Dispositional Abilities.” *Philosopher’s Imprint* 10 (12).
———. 2016. “Ceteris Paribus, I Could Have Done Otherwise.” *Philosophy and Phenomenological Research* 92 (1): 73–85.
———. 2018. “Causation and the Grounds of Freedom.” *Teorema: Revista Internacional de Filosofía* 37 (1): 61–76.
Widerker, David. 1995. “Libertarianism and Frankfurt’s Attack on the Principle of Alternative Possibilities.” *Philosophical Review* 104 (2): 247–61.
Wilson, George M. 1989. *The Intentionality of Human Action*. Stanford University Press.
Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Translated by Anscombe, reprinted in 2009. John Wiley & Sons.
Wolf, Susan. 1990. *Freedom Within Reason*. Oxford University Press.
Wu, Wayne. 2016. “Experts and Deviants: The Story of Agentive Control.” *Philosophy and Phenomenological Research* 93 (1): 101–26.
Zhu, Jing. 2004. “Passive Action and Causalism.” *Philosophical Studies* 119 (3): 295–314.