# Indirect Compatibilism

Andrew J. Latham

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Sydney

31/12/2018

FACULTY OF ARTS AND SOCIAL SCIENCES

BRAIN AND MIND CENTRE

**Declaration**

I certify that the intellectual content of this thesis is the product of my own work. Assistance received in preparing this thesis and sources have been acknowledged.

Andrew James Latham

**Abstract**

In this thesis, I will defend a new kind of compatibilist account of free action, **indirect conscious control compatibilism** (or indirect compatibilism for short), and argue that some of our actions are free according to it. My argument has three components, and involves the development of a brand new tool for experimental philosophy, and the use of cognitive neuroscience. The first component of the argument shows that compatibilism (of some kind) is a conceptual truth. Contrary to the current orthodoxy in the free will literature, which is that our concept of free will is an **incompatibilist concept** - a concept according to which we have free will only if determinism is false - I will show that our concept of free will is in fact a **compatibilist concept** - a concept according to which we can have free will even if determinism is true - and I do so using a new experimental philosophy methodology inspired by two-dimensional semantics.

Of course, even if our concept of free will is a compatibilist concept, this does not mean that there are any free actions in the world: the current empirical evidence from the brain sciences appears to show that there might be no, or very few, free actions in the world, even on many compatibilist understandings of what it would take for there to be free will. The second component of the argument addresses this concern by extending our understanding of compatibilism. Agents act freely either when their actions are caused by compatibilistically acceptable psychological processes, or are indirectly caused by those same processes. Hence the name of my account: indirect compatibilism.

The final component of the argument defends my new account against some interesting objections and provides evidence from cognitive neuroscience that some of our actions count as free by the lights of indirect compatibilism.

**Keywords**: free action; free will; compatibilism; experimental philosophy; cognitive neuroscience

For Norma Latham

**Acknowledgements**

I am heavily indebted to many people without whose support this thesis would not be possible. First, I would like to thank my supervisor Professor David Braddon-Mitchell and associate supervisor Professor Bernard Balleine. Associate Professor Kristie Miller has put in a huge amount of work too, for which I am very grateful. I am especially indebted to David and Kristie who had the unenviable task of teaching me academic philosophy from scratch. There enthusiastic support has been a constant boon. I could not have asked for better supervisors to guide me through my PhD candidature.

Second, I would like to thank my colleagues. In particular, Dr Lok-Chi Chan, Dr James Norton, and Dr Michael Duncan who provided me with much needed in-depth feedback and friendship over my PhD candidature. I would also like to thank the members of Kristie's weekly group meetings: those discussions have exposed me to and educated me on a variety of metaphysical positions and topics in analytic philosophy.

Third, I would like my partner Katherine 'Katie' Round who has stood by and supported me throughout my PhD candidature. I would not have made it to the end without you. I would also like to thank my parents Bud and Margaret Latham who have consistently supported and encouraged my academic pursuits. It is now over!

Fourth, I would like to thank everyone who has contributed to my research across my PhD candidature. This includes (but is not limited to) those who have attended, and provided feedback on, talks I have given on this material. Those who have reviewed, and provided feedback on papers I have submitted on this material. Panel members during my annual performance reviews, and so on. Further, I would also like to express my (pre-emptive) gratitude to my thesis examiners for agreeing to read and assess this PhD thesis.

This thesis was partially completed under an Australian Postgraduate Award (APA) and University of Sydney Merit Award.

## Contents

## Chapter 1
## Introduction

In this thesis, I will defend a new compatibilist account of free action: **indirect conscious control compatibilism**. Henceforth, I will refer to this view as **indirect compatibilism**. It is a new, compatibilist account of free action that is the combination of two theses. The first is that the best understanding of the conceptual relationship between determinism and free will is that it is a **conditional concept** - roughly, that indeterminism or libertarian powers are necessary if they are actual, but not if they are not. The second is **indirection** – roughly, that actions are free either when they are caused by standard conscious deliberative processes, or else by sub-personal level processes influenced in various ways by conscious deliberative processes.

One way to understand indirect compatibilism is as an extension and modification of one of the simplest and most basic thoughts about free action: that we act freely when and only when we consciously choose actions on the basis of our desires, in the ordinary way, and then we perform those actions that we choose, and there is no objectionable defeater of an alarming kind.[1] While this is not my view - the 'only when' clause in particular is something this thesis modifies - it does inherit the basic intuition that choice of this kind is sufficient for there being free will, and that those choices play some role in there being free actions even where they are not necessary.

Interestingly, such a view has rarely been explicitly defended – and I do not plan to defend it explicitly at any length either. Instead this thesis primarily proceeds by considering two important impediments to this idea.

The first is that the 'ordinary way' in which we consciously choose actions on the basis of our desires is either deterministic or merely chancy, and it is a necessary condition for the will to be free that this process is neither deterministic nor merely chancy. That is, for there to be free will, these conscious choices must be *indeterministic*, whilst not being *merely chancy*.

I will attempt to overcome this impediment partly by appealing to experimental philosophy (as I will foreshadow in greater detail shortly), which shows that, in a subtle and interesting way, the folk concept is friendly to compatibilism. Of course, there are other ways to proceed here: one could agree that the folk concept is incompatibilist (as much of the previous evidence seemed to suggest) but suggest conceptual reform of some kind. But even if one proceeded that way, it is important that the reformed concept be recognisably a concept of free

[1]Example defeaters might include mind control, neurological intervention, the wrong kind of advertising, and so on.

will, and my work here suggests that such reformed concepts can indeed be concepts of free will, while explaining why it might seem otherwise.

The second impediment is the suggestion that as a matter of fact few actions *are* under the direct control of conscious processes on the basis of our desires, and thus that few actions are free and, perhaps, thus that according to this simple and basic account we do not have free will. My solution here will be to suggest a way in which many such actions might be indirectly under the control of these conscious processes, and that it is (modulo defeaters perhaps) necessary and sufficient for free action that the action be either directly or indirectly under conscious control.

So much for the modifications: but what of the basic idea? Why am I addressing objections to a theory and modifying it in light of them, without specifically defending it? Why do I regard it as any kind of default?

One thing to say, and perhaps my official reply, is that this dissertation is a defence of a conditional claim: that if you find such a simple account pre-theoretically attractive, there are modifications and responses which allow it to survive the most powerful objections.

But there is a little more to be said. Much of the free will debate seems to accept, more or less, that something like this default account is appealing, because much of the literature is concerned with objections to that default account, or exploring the consequences of it.

Let me illustrate this with a few examples. First: Galen Strawson. Strawson gives an argument that no-one ultimately possesses moral responsibility. What are the assumptions of this argument? In part that conscious control is a necessary condition of freedom of the will:

> "[...] it is not merely that one must have caused oneself to be the way one is, mentally speaking. One must have consciously and explicitly chosen to be the way one is, mentally speaking, in certain respects, and one must have succeeded in bringing it about that one is that way." (pg. 6).

Strawson makes trouble for the idea of responsibility by requiring this to iterate: he thinks that you need to have conscious control of the psychological features which are the prerequisites for conscious control as well, and obviously this iteration is going to fail. I do not think that the default conscious control account is committed to the view that we need conscious control of our conscious control. Instead, it's just the view that conscious control plays the role that we suppose it does in the ordinary cases that we think of as free. But that need not detain us here. The thought is just that this conscious control requirement seems like an obvious one to start with - something that is either obviously true and will cause trouble, or is an obvious starting point, to be revised only when required by that trouble.

Consider another example. Peter Van Inwagen thinks that the thesis that we have free will is "the thesis that we are sometimes in the following position with respect to a contemplated future act: we simultaneously have both the following abilities: the ability to perform the act and the ability to refrain from performing that act." (2008; pg. 330). I think that what Van Inwagen has in mind with respect to *contemplating future acts* is the default conscious control account. Consider the following excerpt from his seminal paper *The Incompatibility of Free Will and Determinism* (1975) that describes a judge who chooses not to raise their hand in order to spare a criminal's life (he very rarely goes into much detail on the psychological processes that surround our free actions, so savour it):

> "We may also suppose that the judge was unbound, uninjured, and free from paralysis; that he decided not to raise his hand at T only after a period of calm, rational, and relevant deliberation; that he had not been subjected to any 'pressure' to decide one way or the other about the criminal's death; that he was not under the influence of drugs, hypnosis, or anything of that sort; and finally that there was no element in his deliberations that would have been of any special interest to a student of abnormal psychology." (pg. 190-191).

Recall that according to the default conscious control account, an agent acts freely only if their action is caused by their conscious deliberation and standard decision making processes. According to the default account, the judge clearly acts freely. The judge's action was caused by "a period of calm, rational and relevant deliberation". Further there was nothing abnormal about the judge's decision making processes that brought about his action: "there was no element in his deliberations that would have been of any special interest to a student of abnormal psychology". However, Van Inwagen does not think that the judge acts freely in a deterministic world. In a deterministic world containing only deterministically governed conscious deliberations and standard decision making processes, the judge does not act freely. So what appears to be going on is that Van Inwagen is taking this 'calm rational and relevant deliberation' - surely something that entails that there is conscious control, with perhaps some *extra* rationality requirements – to be both necessary for free will, and also *prima facie* sufficient, such that it comes as worthwhile news to note that, given the possibility of determinism, it is not sufficient.

One of the many seminal contributions made by Harry Frankfurt to the free will literature was the establishment of a new kind of compatibilist account of free will: so-called higher-order accounts of free will. This account was first articulated in *Freedom of the Will and the Concept of a Person* (1971). I think this intervention, too, makes most sense understood as a surprising response to the default conception (exactly how I think it connects with my account

will come later). The key insight of these accounts is that what it is to be free is not a question that concerns the connection between conscious deliberation and standard decision making processes, and actions, rather, questions of freedom just concern the connections within our own psychologies. As Frankfurt writes:

> "The question of the freedom of his will does not concern the relationship between what he does and what he wants to do. Rather, it concerns his desires themselves." (pg. 15).

But what is the relevant connection between "what he does and what he wants to do" that we might have originally thought questions about freedom were concerned about? Here I think Frankfurt has something like the default conscious control account in mind. As he writes:

> "Human beings are not alone in having desires and motives, or in making choices. They share these things with the members of certain other species, some of whom even appear to engage in deliberation and to make decisions based on prior thought." (pg. 6).

He elaborates later when describing a 'wanton', someone who lacks any higher-order desires. He writes:

> "The fact that a wanton has no second-order volitions does not mean that each of his first-order desires is translated heedlessly and at once into action. He may have no opportunity to act in accordance with some of his desires. Moreover, the translation of his desires into action may be delayed or precluded either by conflicting desires of the first order or by the intervention of deliberation. For a wanton may possess and employ rational faculties of a higher order. Nothing in the concept of a wanton implies that he cannot reason or that he cannot deliberate concerning how to do what he wants to do." (pg. 11).

According to the default conscious control account, someone acts freely only if their action is caused by their conscious deliberation and standard decision making processes. The wanton in this example has desires, and, what's more, is able to reason and deliberate concerning how to do what he wants to do. Thus according to the default account the wanton acts freely.[2] The surprise and interest in Frankfurt's and other higher-order theorists view comes from seeing a psychology which seems to satisfy the default conception and yet allegedly lacks freedom and autonomy altogether. That's because, according to Frankfurt, free will is not about the connection between psychological processes and actions as the default conscious control account tells us. Instead, it

[2]I think that the reason we are taken aback by this conclusion is not for the reason that Frankfurt and other higher-order theorists think. Higher-order desires seem ubiquitous to us and are a common feature in our own conscious deliberations and standard decision making processes. Thus while wantons might act freely according to the default conscious control account of free action they are deeply psychologically impoverished relative to you and I.

is to be found in the structure of the psychological processes themselves, and if those structures are lacking then there are no free actions.

Finally, in the cognitive neuroscience literature about free will it is especially clear that most have something like the default conscious control account in mind. For example, Haggard, et al. (2002) write:

> "Humans have the conscious experience of "free will": we feel we can generate our actions, and thus affect our environment [...] Normal human experience consists of a coherent stream of sensorimotor events, in which we formulate intentions to act and then move our bodies to produce a desired effect." (pg. 382).

Similarly, Bode, et al. (2014) write:

> "The belief that we have free will - that we are the authors of our own fate - is fundamental tor our self-concept and our identity as human beings. Our intuition tells us [...] we can reflect on our reasons, deliberate on our options, and consciously choose to do *otherwise* if we wish." (pg. 636).

What concerns brain scientists is whether our default conscious control account is answered to by anything in the world. More specifically, is there any evidence in the brain sciences that can vindicate the idea that our conscious experience of reflecting on reasons, deliberating on options, formulating intentions, and consciously choosing, is causally efficacious in the generation of our actions. Surprisingly there is accumulating, albeit contentious, evidence that the opposite is true: our conscious deliberations and standard decision making processes do not cause actions. And if that is right then none of our actions are free.

Of course, the fact that these authors, and many others, have taken it to be clear that what I am calling the 'default view' is an obvious starting point does not make the view true, nor even make true that it is the correct starting point. But it does make it something worth exploring. This thesis will respond to the most powerful objections to the basic idea behind the default view, and make some suggestions about how to think about others. If the reader accepts that these objections are dealt with, she will then be in a better position to decide if the whole package I propose is acceptable to her.

Just how crucial is the conscious control requirement to the upshot of this thesis? The first half of the thesis does not depend on it in any way. The first half establishes that a certain, unusual, version of compatibilism appears to hold of the folk concept of free action. That will be of interest to readers who have little interest in the conscious control thesis, and accept versions of compatibilism which do not rely on that thesis. It is, however, necessary to show this for conscious control compatibilism to be vindicated. The second half of the thesis is more directly

concerned with conscious control. But this is not only because I am inclined to accept indirect conscious control as a necessary condition for free action, and hope to show that it can be accepted consistently with holding that enough of our actions are free for us to have free will. It's also because it's a requirement which, as I have shown, is accepted by most researchers in the brain sciences who aim to show that our actions are not free. In criticising their work, it's helpful to accept as many of their assumptions as possible and see if nonetheless most of the actions we intuitively take to be free can be seen to be free. So apart from my own interest in showing that this assumption - which I take to be compelling - is consistent with free will, it also plays a dialectical role in the strongest criticism of these authors.

So, in this thesis, I will defend in detail a new compatibilist account of free action centred on conscious control, against two profound charges: that our concept of free action is incompatibilist, and that our conscious deliberations and standard decision making processes cause no, or very few, actions. I will begin by establishing that our concept of free action is compatibilist, using a new experimental philosophy methodology inspired by two-dimensional semantics. Then I will show that we can address the concerns of the brain sciences by extending our understanding of which actions are under our conscious control. Agents act freely either when their actions are caused by conscious deliberations and standard decision making processes *or* when their actions are indirectly caused by these same processes. Hence the name of my account: *indirect compatibilism*. At the end of this thesis, I will defend indirect compatibilism against some interesting objections and provide evidence from cognitive neuroscience that some of our actions count as free according to this new account.

### 1.1. Freedom-conferring mechanisms

In order to introduce the topic of free action from the perspective of a conscious control account, we do best to begin with an example. For now, I will bracket the relevance of determinism to free will. While all these psychological processes might be determined I will assume compatibilism is true for the moment. So, consider the following paradigm case of someone who acts freely. Katie is deliberating about whether to pursue a PhD in Philosophy at the University of Sydney or a PhD in Psychology at the University of Auckland. Before making her decision, she consciously deliberates. Currently living in Auckland, she weighs up the costs of uprooting and moving away from her friends, family and partner, against the benefits to her education, job prospects and passion for knowledge and learning. She simulates an assortment of various possible outcomes for each choice. Eventually, she chooses one option or the other, and through that choice adopts one class of actions over the other. For example, if Katie chooses to

pursue a PhD in Sydney she performs the set of actions that will bring about her choice to study there. By doing so, she refrains from performing the alternative set of actions that would have led her to study in Auckland.

To get clearer on what I mean to be the rough intuitive target for what it takes to rightly say someone 'acts freely', we can contrast Katie's case with a paradigm case of someone who acts unfreely. At a medical check-up, Mila's doctor tests her patella-reflex. When Mila is struck just below the knee-cap with a little rubber hammer, her leg jerks upwards. The leg-movement is automatic. Mila does not choose to move her leg. Exhibiting the patella-reflex is simply a consequence of the physiology of bipedal beings like us, which allows us to maintain balance with little to no conscious thought or control.

Generally speaking, whenever we act freely it's because our actions are caused by certain psychological processes and whenever we act unfreely it's because those psychological processes that cause free actions are absent. For example, the reason Katie acts freely when she decides to pursue a PhD in Sydney is because her actions were presumably caused by her conscious deliberations, her desire to study at Sydney, her reasons to study at Sydney, and so on. Conversely, the reason Mila's action is unfree is because her action was not caused by any psychological processes of the kind that caused Katie's action. Mila's reflex was not caused by the conclusion of any conscious deliberation, nor was it caused by any of her desires or reasons for action, or any psychological processes at all for that matter. Instead, Mila's action was simply the result of being struck below the patella in conjunction with some facts about her physiology and the physiology of beings like us. When we act freely then, we act for our reasons, to fulfil our desires, or after having mentally simulated various course of actions and their projected outcomes, and so on. Perhaps we act freely, as a result of all or some combination of these psychological processes. Of course, the factors listed here do not exhaust all the differences between the two examples. Nor am I taking any specific psychological processes to be necessary or sufficient for free action at this moment in the thesis.

One way we can get clearer on the difference between cases where people act freely and cases where people act unfreely is by appealing to the personal and sub-personal distinction first introduced by Daniel Dennett in *Content and Consciousness* (1968). Dennett distinguishes between the "explanatory level of people and their sensations and activities" and "the sub-personal level of brains and events in the nervous system". Applied to the case of free action, we might think that people act freely only if their actions are caused by certain personal-level psychological processes. So when Katie reports her reasons, or her desires, or the outcome of her deliberation, she acts freely only if her reasons, desires, or conscious deliberation, and so on, played the role

she supposed they did in causing her actions. Conversely, the reason Mila's action is unfree is because there are no personal-level explanations for her action. Instead, Mila's action was caused by some sub-personal level mechanism. I will call the conscious psychological processes (along with their neural correlates)[3] that cause free actions: **freedom-conferring mechanisms**. One of the major issues I will address in this thesis is whether any of our actions are caused by freedom-conferring mechanisms, or whether they are, as some brain scientists claim, all caused by sub-personal level mechanisms. Briefly, my answer will be that while I think some of our actions are caused directly (in a sense which will be made more precise later) by freedom-conferring mechanisms, the number that are is likely far fewer than we might have originally thought. For the time being though, let's set this worry aside.

Why should we care about freedom-conferring mechanisms? One reason is because the notion captures the vast majority of compatibilistically acceptable psychological processes free will theorists think, or might think, are crucial for free actions (for an up to date review see McKenna & Coates, 2015; O'Connor & Franklin, 2018).[4] If all our actions are caused by sub-personal level mechanisms, which are not themselves under the control of compatibilistically acceptable psychological processes along with their neural correlates, then almost every free will theorist is going to think there are no free actions (see Mele 2010). Of course, according to different accounts of free will some psychological processes are more crucial than others. For example, the idea that actions caused by our desires are free has its roots as far back as Hobbes (1651/1997) and Hume (1740/1978). Further, according to Fischer and Ravizza's (1998; though see also Fischer, 2004) popular reasons-responsiveness account, an agent acts freely when they recognize and react in a flexible manner towards reasons for action. In this thesis, I will make no attempt to adjudicate which psychological processes are crucial, or most important, for free action, though I am partial to desire-based accounts. Instead, one of the major tasks of this thesis is to argue that actions indirectly caused by freedom-conferring mechanisms (again, something I will spell out more shortly), in some circumstances, are free as well as are those directly caused by those mechanisms. However, if you have prior theoretical commitments in this area, whenever you read 'freedom-conferring mechanism', feel free to read-in your preferred compatibilistically acceptable psychological process, or processes.

[3]Some possible accounts of the personal-level include global neuronal workspace theory (e.g., Dehaene, Kerszberg, & Changeux, 1998; Dehaene & Naccache, 2001; Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006; though see also Baars, 1988), recurrent-processing theories (e.g., Lamme, 2006; 2010), higher-order representational theories (e.g., Kriegel, 2003; Carruthers, 2005; Rosenthal, 2005) and information integration theory (e.g., Tononi, 2004; 2008; though see also Latham, Ellis, Chan, Braddon-Mitchell, 2017). While I prefer global neuronal workspace theory, nothing that I say in this thesis hangs on this.

[4]With any attempt to introduce broad categorizations in philosophy there are always exceptions. For example, see Ned Markosian's agent-causal compatibilism (1999; 2012).

Another reason to care about freedom-conferring mechanisms is because the folk also care about these mechanisms with respect to free action. For example, Daniel Shepherd (2012; 2015) observed that people only straightforwardly judge that an agent acts freely when the agent's conscious psychological processes were central to causing their actions and controlling their behaviour. Furthermore, when brain scientists attempt to show that people do not act freely, their arguments are informed by results that purport to show that the real causes of our actions are sub-personal level mechanisms, not, as we might have supposed, freedom-conferring mechanisms. Once again, I will return to introduce the challenge posed by the brain sciences in more detail in §1.3. However, before I do that I want to introduce and outline my response to the defining problem of the free will literature: the problem of determinism.

### 1.2. Conditionalism and the problem of determinism

Until now, I have bracketed off the problem of determinism for free will. Determinism is the thesis that any earlier state of the universe, in conjunction with the laws of nature, is logically sufficient for any later state of the universe. Free will theorists who think that free will is compatible with determinism being true are referred to as compatibilists. On the other hand, free will theorists who think that free will is incompatible with the truth of determinism are referred to as incompatibilists, and incompatibilists who think that we have free will are referred to as libertarians.[5] Many libertarians have been convinced by arguments, such as Peter van Inwagen's *Consequence argument*, (1983), that purports to show that if determinism is true then there are sufficient causes for our behaviour that are outside our control, and if that's true then we could not have done anything other than what we actually did, so we are not free.[6] For example, recall the case of Katie choosing between the University of Sydney and the University of Auckland. If determinism is true, then it was always the case that Katie would go to the University of Sydney.

[5]Libertarians can be further classified according to the kinds of powers they think agents exercise to overcome indeterminacies. *Event-causal* theorists think there are certain reducible psychological processes that can overcome indeterminacies. Perhaps the most widely known account of this type comes from Robert Kane in *The Significance of Free Will* (1996). On Kane's account indeterminacies are present inside the agent when they are deciding between competing visions about what to do or who they should become. These indeterminacies, stirred-up and magnified by chaos in the brain, impede either vision from being actualized. Only through the agent's 'efforts' is the indeterminacy overcome resulting in one of their actions being actualized. However, I find it difficult to grasp how the psychological effort of an agent can overcome genuine chance in the world.

*Agent-causal* theories posit the strong emergence of fundamental agent-causal powers (see for example, Clarke 1993; 1996; O'Connor, 1995; 2000). I tend to think that these views properly understood either collapse into event-causal libertarianism, or are dualistic. This is not to say that these positions are necessarily false. For example, some such as Derek Pereboom (2001) think that agent-causal powers are necessary by the lights of our concept of free will. However, their unintelligibility and drastic departure from current scientific understanding that leads us to be error theorists about free will.

[6]In his own words van Inwagen (1983) summarises the Consequence argument as follows: "If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us." (pg. 5). See also Strawson's (1994) *Basic argument*.

Libertarian free will requires that Katie could have gone to the University of Auckland, even given with the same past and the same laws of nature.

The idea that free will is incompatible with determinism is not an idiosyncratic worry of some free will theorists. The orthodox view in the literature is that our folk concept of free will is incompatibilist. This is supported by some excellent evidence from experimental philosophy (e.g., Nichols, 2004; Nichols and Knobe, 2007) and the numerous self-reports provided by free will theorists in the literature. For example, Robert Kane (2005) writes:

> "In my experience, most persons resist the idea that free will and determinism might be compatible when they first encounter it. The idea that determinism might be compatible with free will looks at first like a "quagmire of evasion," as William James called it, or a "wretched subterfuge" as Kant called the compatibilism of Hobbes and Hume. If compatibilism is to be taken serious by ordinary persons, they have to be talked out of this natural belief in the incompatibilism of free will and determinism by means of philosophical arguments." (pg. 12-13).

It is commonly thought that our concept of free will requires that indeterminism be true in order for any of our actions to be free. Hence in order for me to be free there must be at least one action where I could have acted otherwise, given the same past and the same laws of nature.[7] This is to think that our folk concept of free will is incompatibilist. In Chapter 2 of this thesis I will argue that the current orthodoxy is wrong about the folk concept of free will. The folk concept of free will is not incompatibilist; instead compatibilism of some kind is a conceptual truth.

The main reason I think it is important to show that our concept of free will is compatible with determinism (that is that the folk think it is possible to act freely in a deterministic universe) - or, as I will say, that our concept of free will is compatibilist - is because if our concept is not compatibilist then it will prove a major obstacle to the development of my new account of free action: indirect compatibilism. If the folk do not think it is possible to act freely in a deterministic universe, then it is hard to see how *any* supposed account of free action of a compatibilist kind could deserve to be called 'free action'. By extension if there are no free actions in the world because our world is deterministic then it is hard to see how we could vindicate the idea that we have free will. That is not to say that what is described in this thesis would not be an account of *something*, just that it would not be, as I might have hoped, an

[7]Once again there are exceptions to this broad categorization. For example, Erkstrom (e.g., 2000; 2003) defends a libertarian account of free will that locates indeterminacies at the input of preferences into deliberation. Deliberation itself proceeds in a deterministic fashion. Similar accounts have been described Mele (e.g., 1995; 2006) who refers to the position as *soft libertarianism*, and Dennett (1978).

account of what it takes to act freely. That's because our concepts of free action and free will would be incompatibilist and so indeterminism would be a necessary condition for there being free will.

What is the relationship between our concept of free action and our concept of free will? As I see it our concept of free action is about the aetiology of our actions and our concept of free will is about the general capacity of an agent to reliably act freely. According to indirect compatibilism an action is free only if either it is caused by freedom-conferring mechanisms, or is indirectly caused by freedom-conferring mechanisms. Shortly, I will outline how freedom-conferring mechanisms can indirectly cause free actions. Briefly for now though, I will say that any sub-personal level mechanism influenced[8] by a freedom-conferring mechanism causes free actions, albeit indirectly. An agent has free will then only if their actions are reliably caused by their freedom-conferring mechanisms, or sub-personal level mechanisms influenced by freedom-conferring mechanisms. In this thesis I will not attempt to say how many of our actions have to be free in order for us to have free will. However, at the very least I think that the number of actual free actions will need to roughly align with the number of actions that we typically and pre-theoretically judge to be free, which I tend to think, along with the folk, is a significant proportion of our actions.

It's important to note that one major strategy to downplay the importance of findings that the folk conception of free will is incompatibilist, is to argue that while the folk concept of free will is indeed an incompatibilist one, the best concept of free will we can deploy is compatibilist. Most compatibilists, I think, are not engaged in the descriptive project of giving an account of our concept of free will. Instead, they are engaged in the project of conceptual engineering, developing what they think is the best concept of free will for us to deploy (see Manuel Vargas *Building Better Beings* (2013) for an excellent example of this type of project). One of the reasons they are engaged in this project is because they think there is something deeply problematic with libertarian free will. For example, many compatibilists have been convinced by arguments such as the *Luck argument* (e.g., Mele, 1999; Haji, 2000; Levy, 2011) and others like it. These arguments purport to show that if indeterminism is true and so our actions are undetermined, then it's a matter of chance what we do, and if that's true then we are not free.[9] For example, recall again the case of Katie who is torn between studying at the University of Sydney and the University of Auckland. If indeterminism is true, then whether or not Katie goes to the University of Sydney is a matter of chance. Katie's freedom-conferring mechanisms

[8]Exactly what sort of connections count as influence will be introduced in §1.3 and discussed in Chapter 4.
[9]See also the closely related *Mind* argument (e.g., Hobart, 1934; Nowell-Smith, 1948; Smart, 1961). So-called by van Inwagen (1983) because they were all published in the journal *Mind*.

cannot determine what she will do because if they do then it's no longer true that she could have acted otherwise given the same past and the same laws of nature.

While I think that libertarian accounts of free will are problematic I will not argue to that conclusion in this thesis (though in the midst of a dark night of the soul I think libertarian accounts might be right). Nor, will I argue that we should conceptually engineer and deploy a compatibilist concept of free will because that is the best concept of its kind. Instead, in this thesis I will take a different route and argue that the folk concept of free will is already a compatibilist one.

In Chapter 2, I will argue for an unorthodox view in the literature, according to which our folk concept of free will is a compatibilist concept, albeit of a unique kind. The folk concept of free will is a *conditional concept*. Briefly, a conditional concept is a single concept that has an indexical component which fixes what is necessary for the concept to be satisfied, depending on what the actual world is like (Braddon-Mitchell, 2003). So roughly, if the actual world is indeterministic, then indeterminism is necessary for free will. However, if the actual world is deterministic, then indeterminism is not necessary for free will. The reason the conditional concept is a compatibilist one is because once you have fixed the personal-level processes you think are necessary for free will, then no matter how things turn out to be with respect to determinism, you will judge that we sometimes act freely. In this chapter I will also present evidence from my own experimental philosophy investigation that shows the folk concept of free will has a conditional structure, and thus shows the folk are free will compatibilists.

### 1.3. Indirection and the problem of the brain sciences

One of the ways we act freely, according to indirect compatibilism, is when our actions are directly caused by freedom-conferring mechanisms. I will call the view that an action is free only if it is directly caused by freedom-conferring mechanisms: **direct conscious control compatibilism**. Henceforth, I will refer to that view as **direct compatibilism**. However, even after I have established that these compatibilistically acceptable psychological processes are acceptable as causes of our free actions, it could still turn out that there are no free actions, and thus no one in the world has free will.

There is accumulating, albeit contentious, evidence that freedom-conferring mechanisms cannot be playing the role we suppose they do in causing actions. Consider the following basic paradigm. Participants are asked to watch a time-keeping device and report the time shown when they first feel the urge to perform some simple motor action, such as a wrist flex, or cognitive action, such as basic arithmetic. Shortly after they have performed the simple action, the time-

keeping device stops and the experimental task is reset. It's robustly found that the onset of neural activity associated with an action occurs well before participant's first report their urge to perform the action (e.g., Libet, Gleason, Wright, & Pearl, 1983; Bode, He, Soon, Trampel, Turner, & Haynes, 2011). On the basis of these results, some brain scientists conclude that our freedom-conferring mechanisms cannot be the cause of our actions (e.g., Gazzaniga, 2011; Greene & Cohen, 2004; Libet, 1999). After all, if freedom-conferring mechanisms were causing actions, then the time participant's first report their urge to perform an action should occur before (or at least roughly in line with) the onset of supposed neural activity for that action.

In this thesis, I will not argue that there is anything wrong with the methods and results of these studies, even though there are a number of good critiques already in the literature that are certainly worth developing. For example, you might wonder whether people can reliably and accurately report the timing of their urge to perform an action (e.g., Banks & Isham, 2009). On the other hand, you might wonder whether the earliest recorded neural activity is actually related to the action that participants will go on to perform (e.g., Miller, Shepherdson, & Trevena, 2011). In Chapter 3 I will just accept the general findings of these studies and will attempt to develop the strongest argument from the brain sciences against the existence of free actions. Ultimately, I will argue that the conclusion there are no free actions is false. However, there is nearby conclusion that is plausibly true: only a limited number of our actions are directly caused by freedom-conferring mechanisms, and the vast majority of our actions are in fact caused by sub-personal level mechanisms.

Some readers might be satisfied with the result that only a few of our actions are free; but I am less sanguine. That's because if direct compatibilism is right and so only a few of our actions are free, then it's entirely possible that we do not have free will. I will call this the **discrepancy problem**. Recall my earlier description of the relationship between our concepts of free action and free will; while our concept of free action refers to the aetiology of our actions, our concept of free will refers to the general capacity of an agent to act reliably in a free manner. If the only actions that get to count as free are those caused by freedom-conferring mechanisms, then while there might be a few free actions, as we do not act reliably in a free manner, so we do not have free will.

However, I don't think that free actions are only those directly caused by freedom-conferring mechanisms. Instead, I think that some sub-personal level mechanisms cause free actions as well. But why should we think that? Here is a suggestive thought. I find it very striking that brain researchers ever thought that the simple, repetitive actions in these experiments were free. Yet they must, since otherwise, why would they think that their results speak to the issue of

free action? I do not think that the judgement these simple, repetitive actions are free is wrong; instead I think this judgment is right and reveals something interesting about our concept of free action. In these experiments people were asked to repeatedly perform a very simple action for hundreds of trials. Once people had freely decided to participate, most people probably automated their performance, so that their actions were not being caused by freedom-conferring mechanisms, but instead were being caused by a sub-personal level mechanism. Insofar as we think that we correctly judge that the actions performed during these experiments are free, then *prima facie* it seems as though some sub-personal level mechanisms cause free actions.

Consider again the earlier case of Katie choosing between studying at the University of Sydney and the University of Auckland. Katie chose to study at the University of Sydney and as a result 'performs the set of actions that will bring about her choice to study there'. The choice of language in the original case was intentional and what I am particularly interested in, in this thesis, is the status of the actions that bring about Katie's choice to study at the University of Sydney. Once Katie has freely chosen the University of Sydney, she might automate most of her actions to bring about her choice to study there. That is, some of her future actions might be caused by a sub-personal level mechanism whose function it is to bring about her choice to study at the University of Sydney. In this thesis I will argue that even if only a few of our actions are directly caused by freedom-conferring mechanisms, most of our actions might still be free if they are caused by a certain kind of sub-personal level mechanism: those influenced by our freedom-conferring mechanisms.

In Chapter 3 I will develop the notion of an **action procedure**. An action procedure is a sub-personal level mechanism that causes actions, or influences other action procedures, in a contained and reliable fashion. So for example, we can imagine a simple hand-waving action procedure that causes a hand-waving action. The action procedure is contained in the sense that it will only cause actions of the hand-waving kind (contained) and will rarely fail to cause hand-waving actions when activated (reliable). An action procedure that is influenced by a freedom-conferring mechanism is a **controlled action procedure**. According to indirect compatibilism, the actions caused by controlled action procedures are free, albeit indirectly, because they are the result of a sub-personal level mechanism influenced by the agent's freedom-conferring mechanisms. So for example, the reason we judge that the participants of these experiments act freely is because when they freely decided to participate in the experiment they might have created an action procedure that causes the appropriate actions in response to some experimental stimuli. Further, we can imagine that Katie's free choice to study at the University

of Sydney might have resulted in a much more complex action procedure which causes some of the actions that bring about her choice to study there.

In Chapter 4 I will go into detail about how people might influence action procedures. For example, one-way people might influence action procedures is by creating them. But this is not the only route to influencing action procedures. For example, people might have the capacity to destroy an action procedure that they possess. They might trigger an action procedure to come on. They might monitor an action procedure and turn it off, if they so choose. Further, people might update an action procedure they already have, to perform a different function. In each case, the resulting controlled action procedure causes free actions because it is influenced by the agent's freedom-conferring mechanisms. In Chapter 4 I will also argue that indirect compatibilism is a better account of our concept of free action than is direct compatibilism because it is better aligned with our pre-theoretic free action judgments (our judgments about which actions are free and unfree) and vindicates the thought that we have free will.

According to indirect compatibilism, then, an action is free only if it is directly caused by a freedom-conferring mechanism, or indirectly caused by a freedom-conferring mechanism via a controlled action procedure. In Chapter 5 I will address a number of interesting objections to my account of indirect compatibilism. For example, do controlled action procedures that we create but can no longer intercede on, cause free actions? Are indirectly free actions caused by controlled action procedures just as free as those directly caused by freedom-conferring mechanisms? I will also offer up one piece of evidence from my own investigations in cognitive neuroscience that I think shows people updating their action procedures, thus showing that a necessary condition of my account, indirect compatibilism is fulfilled.

### 1.4. Indirect compatibilism and autonomy

There is a class of theories in the autonomy literature that are often viewed as compatibilist accounts of our free action and free will. When they are seen in the free will literature they are broadly characterized as higher-order accounts of free will (e.g., Frankfurt, 1971; 1987). According to these accounts, an agent's action is free if and only if the action is in accordance with the will the agent wants. For example, imagine Luca who desires a chocolate cookie and so goes and obtains a cookie. Luca acts freely, on such an account, if and only if she desires her desire for the cookie to be a part of her will. One the other hand, if Luca desires the chocolate cookie, but desires that she did not have her desire for a chocolate cookie, then she doesn't act freely when she eats the cookie. For ease of explication I will just talk in terms of desires, but what kind of psychological states or processes people identify with for the purposes of acting

freely varies according to the account in question. Some other examples include identifying with what you love (e.g., Frankfurt, 1994; Shoemaker, 2003) your plans (e.g., Bratman, 1997; 2007) and your perceptions of the good (e.g., Watson, 1975; 1987).

The main reason I raise these accounts here is because there is a conflict in the judgements issued by these accounts, taken as an account of our concept of free action, and the account of free action that I will be offering in this thesis: indirect compatibilism. Consider again the case of Luca and the cookie. According to my account if Luca reports a standard desire for the chocolate cookie as what caused her to eat the chocolate cookie, then provided her desire is what caused her to eat the cookie, then she acted freely. That's because insofar as our desires are one kind of freedom-conferring mechanism, provided Luca's freedom-conferring mechanism caused her action, then she acts freely. However, this judgment would be wrong according to a higher-order account. Given that Luca also has the desire that her desire for the cookie is not a part of her will, then Luca's act of eating the cookie is not free. That's because Luca does not identify with the desire that causes her action in the right sort of way demanded by these accounts.

But what to say about the conflict between these two accounts? Here are two things you might say. First, higher-order accounts describe yet another necessary condition for free action. It is not enough for an action to be caused by a freedom-conferring mechanism; the freedom-conferring mechanism must be one that we identify with in the right kind of way for it to cause free actions. Second, higher-order accounts of free action are best not thought of as account of free action but instead as an account of something over and above free action: *autonomy*. Thoughts about autonomy, while closely related to thoughts about free action, are nonetheless orthogonal. Actions can be either free or unfree as well as autonomous or non-autonomous. I think that my preference is for the second option, but I will not argue for this in the thesis. What is important going forward is that when I talk about free action I am referring to those actions caused directly and indirectly by freedom-conferring mechanisms and not those that we might think of as autonomous actions.

### 1.5. The long free road ahead

In summary then, here is the plan for the thesis. In Chapter 2 I face up to the classical problem of free will. In order successfully to defend the idea that actions are free when they are caused directly or indirectly by freedom-conferring mechanisms, some version of compatibilism must be an acceptable account of free action. In Chapter 2 I will argue that the orthodox view, according to which the folk concept of free will is incompatibilist, is mistaken. The folk concept of free will

is a unique kind of compatibilist concept: a conditional concept. What makes the conditional concept a compatibilist concept is that when you hold fixed whatever personal-level processes you think are necessary for free action, then there is no discovery that you could make with respect to determinism, which will lead you to conclude that you are unfree. I will support the claim that the folk concept of free will is a conditional concept and thus the folk are free will compatibilists, with some positive evidence from my own experimental philosophy investigations.

In Chapter 3 I will address two challenges posed by the brain sciences to free actions and free will. The first challenge is the **broad challenge**: evidence from the brain sciences shows that our freedom-conferring mechanisms cause no actions, and so none of our actions are free. The second challenge is the **narrow challenge**: evidence from the brain sciences shows that our freedom-conferring mechanisms cause very few actions. I will argue that while the *broad challenge* is wrong, the *narrow challenge* is plausibly right and no less problematic. That is because if our freedom-conferring mechanisms cause very few of our actions, then far fewer actions than we initially supposed will be free. Further, if not enough of our actions are free, then we do not have free will: the *discrepancy problem*. I will argue that we can accommodate many of our free action judgments by acknowledging that most cases that we judge to be free are not free because they are caused directly by freedom conferring mechanisms, but because they are indirectly caused via action procedures which those freedom-conferring mechanisms influenced in some way.

In Chapter 4 I will elaborate on what it means to be an indirectly free action. Remember that according to indirect compatibilism an action is free only if it is directly caused by freedom-conferring mechanisms, or indirectly caused by freedom-conferring mechanisms via controlled action procedures. This means that any sub-personal level mechanism that was created, or can be destroyed, triggered, monitored, or updated by freedom-conferring mechanisms, cause free actions. In Chapter 4 I will also argue that indirect compatibilism is a better account of our concept of free action than straightforward compatibilism because it better accommodates people's pre-theoretic judgments about which actions are free (and unfree), and vindicates the thought that we have free will.

Finally, in Chapter 5 I work through what I think are some of the most interesting objections to indirect compatibilism. For example, this includes working through potential cases that my view says are free but *prima facie* look unfree. Further, some people might have the intuition that there are substantive differences between free actions caused by freedom-conferring mechanisms and those caused by controlled action procedures: remember, a

controlled action procedure is a sub-personal level mechanism that is influenced by a freedom-conferring mechanism.

In Chapter 5 I will also highlight one finding from my own cognitive neuroscience investigations that I think is evidence of people updating an action procedure, and thus shows that an important empirical *necessary* condition is satisfied by my view. The full paper of this result can be found in Appendix A. The thesis now starts right at the beginning at the classical problem of free will. As I take my account to be folk and compatibilist friendly, it had better *not* be the case that the folk conceptually rule out compatibilism from the onset. Let's begin.

## Chapter 2

## Folk free will in two-dimensions[10]

In this thesis I defend the idea that our actions are free when they are either directly caused by compatibilistically acceptable, (but possibly deterministic) processes - i.e. by a freedom-conferring mechanism - or indirectly caused by those same processes - i.e. by a controlled action procedure. However, in order for this is to be so, some version of compatibilism needs to be acceptable as a thesis about free will. In this chapter I will defend that possibility against the following pressing objection: that there is evidence, as a matter of descriptive fact, that our concept of free will is one which doesn't allow deterministically caused actions to count as free.

Is our concept of free will compatible with the thesis of determinism being true? That is, is our concept of free will a *compatibilist concept*? Recall that the thesis of determinism holds that the entirety of particular facts about the past, in conjunction with the laws of nature entails every truth about the future.[11] Compatibilists answer affirmatively; they think that having free will is compatible with determinism being true. According to them, if determinism is true then provided agents have some preferred set of abilities, which vary according to the version of compatibilism at issue, then free actions are those produced by those abilities. For ease of explication I will refer to *whatever the abilities are that when exercised in the production of an action makes that action free according to compatibilism*: **compatibilist powers**. Conversely, incompatibilists take it to be a necessary condition for our having free will that indeterminism is true. That is, they think that having free will is incompatible with determinism being true. I will refer to *whatever the abilities are that when exercised in the production of an action makes that action free according to libertarianism*: **libertarian powers**.

According to the new account of free action that I defend in this thesis – namely, *indirect (conscious control) compatibilism* - actions are free when they are either produced by a freedom-conferring mechanism, or produced in an indirect way via action procedures that were influenced by freedom-conferring mechanisms. Of course, this account is in no way logically connected to compatibilism. We can easily imagine a version of this account that would be true if either our freedom-conferring mechanisms or controlled action procedures were of an incompatibilist kind. But as I develop the proposal in this thesis I will focus solely on

[10]Thanks and credit to David Braddon-Mitchell for his assistance with the arguments and empirical study contained in this chapter.

[11]Of course, there are many different ways of formulating the thesis of determinism and this formulation is from van Inwagen (1983). Nothing I say in this chapter hinges on which formulation you choose.

compatibilist kinds of freedom-conferring mechanisms and controlled action procedures. Hence the name of my account: indirect compatibilism.

A major challenge to developing an account of free action in a compatibilist fashion is that if the folk as a matter of fact don't accept that free actions can exist in a deterministic universe, then it's difficult to argue that any account supposedly of free will of a compatibilist kind deserves to be called 'free will'. One common objection to compatibilist accounts of free will is the 'changing the subject objection'. According to this objection, whenever the compatibilist proposes some account of the compatibilist powers they take to be sufficient for free will, while they are giving an account of something, it isn't an account of free will. That's because our concept of free will is an incompatibilist concept, thus indeterminism is a necessary condition for there to be free will. This objection gains strength from the fact that current orthodoxy within the free will literature is that the folk concept of free will is an incompatibilist concept.[12] The task I have set for myself in this chapter will be to argue that this orthodox view of the folk concept of free will is wrong, and that the folk concept of free will is in fact a compatibilist concept, albeit of a unique kind.

Once again it's worthwhile to note that many compatibilist theorists are not especially concerned with the charge that they are changing the subject. That's because the project they take to have set themselves is not to provide a descriptive account of the folk concept of free will, but rather, to prescribe what our concept of free will ought to be. They argue that the concept of free will we *ought* to possess is not an incompatibilist one, but instead a compatibilist one, irrespective of whatever concept we happen to possess already. As a matter of fact, I agree with them that the best account of free will is a compatibilist account. Further, if I took myself to be engaged in a conceptual engineering project, which I believe that most compatibilist theorists are engaged with, I too would try to show that a compatibilist account of free will best serves the functional role of a free will concept. But I don't think such a project is necessary. That's because I don't think our concept of free will needs to be revised in order to be a

[12]Here are a couple of examples of philosophers explicitly describing the prevalence and tenacity of people's incompatibilist intuitions. For example van Inwagen (1993) writes:

> "It has seemed obvious to most people who have not been exposed (perhaps "subjected would be a better word) to philosophy that free will and determinism are incompatible. It is almost impossible to get beginning students of philosophy to take seriously the idea that there could be such a thing as free will in a deterministic universe. Indeed, people who have not been exposed to philosophy usually understand the word "determinism" (if they know the word at all) to stand for the thesis that there is no free will. And you might think that the incompatibilist of free will and determinism deserves to be obvious – because it is obvious." (pg. 187)

Similar sentiments are shared by Ekstrom (2002), Kane (2005), O'Connor (2000), Perebloom (2001), Pink (2004), Strawson (1986) to name a few. Contra this some theorists such as Eddy Nahmias (2011) think the folk concept of concept is a compatibilist one and that incompatibilist judgments arise out of people misunderstanding the implications of determinism for free will.

compatibilist concept. I think that our concept of free will is already a compatibilist concept to begin with.

One area of research that we might turn to in order to figure out whether or not the folk concept of free will is a compatibilist or incompatibilist concept is experimental philosophy. Existing empirical evidence from experimental philosophy appears to be equivocal. There is some excellent evidence that the majority of the general population have a compatibilist concept of free will, provided by seminal work by Nahmias and colleagues (2005; 2006). But there appears to be equally excellent evidence that the majority of the general population have an incompatibilist concept of free will from studies from Nicholls and colleagues (2007) amongst others. It is my goal in this chapter to explain this apparent inconsistency in the experimental philosophy literature. I will argue that all existing studies in the literature are consistent with the underlying nature of our free will concept being a compatibilist concept, albeit of a particular kind: a so-called 'conditional concept', which I will explain shortly. After I have done this I will then go on to present some of my own empirical data that goes beyond merely showing that the folk concept of free will is consistent with being a kind compatibilist concept, but is strongly suggestive of the fact that the folk concept of free will is such a kind of compatibilist concept.

Here then is the plan for this chapter. First, in §2.1 I will give my account of the conditional concept of free will. I will explain what makes the conditional concept of free will a compatibilist concept and provide some *prima facie* reasons we should think that our concept of free will is a conditional one. After that, in §2.2 I will describe some evidence from experimental philosophy that has either been used to show that the folk concept of free will is a compatibilist concept, or to show that it is an incompatibilist concept. I will explain why the differences in these results are only superficial, and are due to differences in methodological set up, and why in fact they provide weak evidence that we have a conditional concept of free will. Then, in §2.3 I will present data from my own empirical research that provides strong evidence that our underlying concept of free will is a conditional one. In §2.4 I will reply to a series of possible objections. Then finally in §2.5 I will conclude.

### 2.1. What is a conditional concept of free will?

The major task I have set myself in this chapter is to show that the folk concept of free will is a compatibilist concept, albeit of a particular kind: a conditional one. My goal in this subsection is to articulate what a conditional concept of free will is and outline why we might think that the folk concept of free will is a conditional one. I will do this in three parts. I will begin by explaining what a conditional concept is and articulating my conditional analysis of our concept

of free will. Then, I will explain what makes the conditional concept of free will a compatibilist concept and how it can accommodate the tenacity of people's libertarian intuitions. Finally, I will describe some *prima facie* reasons for thinking that the folk concept of free will is a conditional concept. Let's begin.

**2.1.1 Conditional concepts and straightforward concepts of free will**

One route to grasping conditional concepts is as a development of David Lewis's idea of a "best deserver". In 'Should a Materialist Believe in Qualia' (Lewis, 1995) Lewis said that the answer was 'yes and no'. The 'best deserver' of the idea of qualia was something that requires dualism to be true. Of course, as a materialist, Lewis thought dualism was false, but he thought that there was a 'good enough' deserver in the form of a physicalist substitute. Lewis's claim was not that the physicalist substitute answers to our concept of qualia perfectly and that dualists were making a conceptual mistake. Rather, the dualist conception of qualia is indeed, in some sense, the paradigm concept; but the materialist's conception is close enough to count, in some sense, as a concept of qualia. It's easy to see how this thought might be applied to free will. The compatibilist is not claiming that their conception of free will is the uniquely correct conceptual account. Rather, it's close enough to our idea of free will so that if actually libertarian properties are nowhere to be found, but there are things in the world that correspond to the compatibilist conception, they will be 'close enough' to count as free will.[13]

David Braddon-Mitchell (2003) extends this idea using some of the ideas of two-dimensional modal logic by suggesting that there is a single concept with a conditional structure. The idea is that certain concepts of qualia have an indexical component which fixes what is *necessary* for qualia, depending on what the actual world offers up. If the actual world contains certain 'spooky' dualist properties, we will judge not only that they are qualia, but also that they are necessary for it.[14] We would therefore straightforwardly judge that in materialist worlds thought of not as ways actuality might be, but as counterfactual alternatives to it, there is no qualia. On the other hand, if the actual world contains only certain physical states, then they would be judged to be what qualia are actually, and counterfactual worlds containing either those certain physical states or spooky dualist properties would equally count as containing qualia.

Once again, it's easy to see how this idea can be extended to the case of free will, and we might propose a conditional analysis of free will, which proceeds like this:

[13]For those familiar with some of the more nuanced positions in the free will literature *soft libertarianism* (Mele, 1996) could be read as 'best deserver' account. According to soft libertarian the most desirable form of free will has indeterminism is a necessary condition. Whether or not indeterminism is a necessary condition for free will *simpliciter* though is an open question. To the best of my knowledge no one in the free will literature endorses this view.
[14]See also Hawthorne (2002) and Stalnaker (2002).

If the **actual world** is indeterministic, and agents have libertarian powers, then these libertarian powers are what free will is and must be.

Else, if the actual world is deterministic, and agents have their preferred compatibilist powers, then compatibilist powers are what free will is.

Then if the first condition obtains, then counterfactually we would judge that all and only indeterministic worlds containing agents with libertarian powers contain agents with free will. By contrast, if the second condition obtains then counterfactually we will judge compatibilist powers to be sufficient for free will.[15]

To make things even clearer, this conditional analysis of free will can be organized into a simple two-dimensional diagram as follows:

| | | **Possible World** | |
|---|---|---|---|
| | | **I** | **D** |
| **Actual World** | **I** | **T** | **F** |
| | **D** | **T** | **T** |

**'Some agents have free will'**

**Figure 1.** Two dimensional diagram showing the conditional analysis of free will, given the sentence 'some agents have free will'.

Here is how to read the two dimensional table: along the top of the table we see two classes of worlds, indeterministic worlds (I) and deterministic worlds (D). Let's suppose for ease of explication that all indeterministic worlds contain agents with libertarian powers and all deterministic worlds contain agents with compatibilist powers (this assumption can easily be removed with a much more complex diagram). These are 'worlds considered as counterfactual' relative to each other. Down the left hand side, we see the same two classes of worlds, but here they are not thought of as counterfactual alternatives to each other, where one is actual and the other is an alternative. Instead, they are alternatives about how the actual world itself, for all we know *a priori*, might be.

[15]Roskies and Nichols (2008; though see also Björnsson, 2014) suggestively describe a similar conditional analysis of our concept of free will in order to accommodate the results of an experimental philosophy study. I will discuss the results of their study shortly in §2.3. Unfortunately, there results cannot be used to determine whether our free will concept is a conditional one or not.

What we are doing when we read this table, is considering our judgments about whether or not some agents have free will, relative to different contexts (ways things might be, for all we know, only one of which is actual), from the perspective of different indices (ways the actual world might turn out to be). Suppose, then, that the actual world turns out to be indeterministic. From the index of an indeterministic actual world, if we look at counterfactual worlds that are also indeterministic then we will judge that it is true that some agents have free will. This is reflected in the T value in the world at the top left cell of our table. That world is being evaluated from the perspective of an actual indeterministic world (specified on the left of the table). The top right cell contains an F. There, we evaluate what to say about the truth-value of 'some agents have free will' at a deterministic world, from the perspective of an indeterministic actual world. In that case, since we judge that those deterministic worlds do not contain agents with free will, that sentence comes out as false.

On the other hand, suppose that the actual world turns out to be deterministic. Now consider our judgements about 'some agents have free will' at a deterministic counterfactual world (the cell on the bottom right). Since deterministic powers are sufficient for free will, we will judge that the sentence is true in that counterfactual world. Furthermore, since having either compatibilist or libertarian powers is sufficient for having free will conditional on the actual world being deterministic, it follows that we will judge that in any worlds with those powers, regardless of whether they are deterministic or not, agents have free will.[16] Hence 'some agents have free will' will be true when evaluated in counterfactual indeterministic worlds, conditional on the actual world being deterministic. This is reflected in the bottom left cell of the table.

*2.1.1.1 Two dimensional diagrams for a straightforward libertarian concept and straightforward compatibilist concept of free will.* We can also use these two-dimensional tables to get clear about other kinds of free will concepts. Before going on I will give a two dimensional reading of two other rival candidates of the folk concept of free will: the **straightforward libertarian concept** and the **straightforward compatibilist concept**. According to the *straightforward libertarian concept* of free will, libertarian powers are what free will is and must be. And, according to the *straightforward compatibilist concept* of free will, both libertarian and compatibilist powers are sufficient for free will. Unlike the conditional concept of free will, what free will is, and must be, according to the straightforward concepts is not indexed to the way the world is actually.

The reason I am providing a two-dimensional reading of the straightforward libertarian concept and straightforward compatibilist concept of free will is so that I can flag where we should expect to observe differences in judgments between them and the conditional concept of

[16]One objection that I will discuss later in §2.4 is what to say if some compatibilists judge libertarian powers to be metaphysically impossible.

free will. Remember, these are differences in our judgments about whether or not some agents have free will, relative to different contexts, from the perspective of different indices. Where we should expect to find differences in our judgments, depending on which concept we deploy, will become an important matter shortly when discussing current approaches in experimental philosophy to examining our free will concept. Briefly, I don't think that current approaches in experimental philosophy can tell us whether our free will concept is a conditional one or not. That's because with respect to the free will judgments being evaluated by those studies we should not expect any differences between the conditional concept and the straightforward concepts. For now, though, I just want to describe the two dimensional table for the straightforward libertarian concept and straightforward compatibilist concept of free will.

| | | **Possible World** | |
|---|---|---|---|
| | | **I** | **D** |
| **Actual World** | **I** | **T** | **F** |
| | **D** | **T** | **F** |

**'Some agents have free will'**

**Figure 2.** Two dimensional diagram showing the straightforward libertarian analysis of free will, given the sentence 'some agents have free will'.

Let's start with the two dimensional table for the straightforward libertarian concept of free will (see Figure 2). Once again, for ease of explication I will suppose that all indeterministic worlds contain agents with libertarian powers and all deterministic worlds contain agents with compatibilist powers. Imagine that the actual world turns out to be indeterministic (the top row of the table). If we then evaluate the indeterministic worlds (the indeterministic contexts) from the indeterministic world considered as actual (from the indeterministic index) then we will judge that those worlds contain agents with free will. This is reflected in the T value that we find in evaluating 'some agents have free will' in the top left cell. But if we look at deterministic worlds from the index of an indeterministic world, then we will judge that those worlds contain no agents with free will. This is reflected in the F value in the top right hand cell, where we evaluated 'some agents have free will' at a deterministic world, conditional on the actual world being indeterministic.

Now imagine that the actual world turns out to be deterministic (the bottom row of the table). According to the straightforward libertarian concept of free will, compatibilist powers are not sufficient for free will irrespective of what the actual world is like. Thus, no agents have free will actually, nor in any counterfactual worlds that are deterministic. This is reflected in the F value that we find in the bottom right hand cell, in which we evaluate 'some agents have free will' at a counterfactual deterministic world, conditional on the actual world being deterministic. But if we look at indeterministic worlds from the perspective of a deterministic actual world, then we will judge that those indeterministic worlds do contain agents with free will. This is reflected in the T value that we find in the bottom left hand cell, in which we evaluate 'some agents have free will' at a counterfactual deterministic world, conditional on the actual world being deterministic.

It's important to highlight that from the index of the indeterministic actual world there are no differences in free will judgments between the straightforward libertarian concept and the conditional concept of free will (compare Figure 1). Differences in free will judgments between the straightforward libertarian concept and conditional concept of free will are only found when considering people's free will judgments indexed to a deterministic actual world. That's because, according to the conditional concept of free will, compatibilist powers are sufficient for free will when determinism is true actually. Conversely, according to the straightforward libertarian concept of free will, compatibilist powers are never sufficient for free will, even from the perspective of a deterministic actual world.

| | | **Possible World** | |
|---|---|---|---|
| | | **I** | **D** |
| **Actual World** | **I** | **T** | **T** |
| | **D** | **T** | **T** |

**'Some agents have free will'**

**Figure 3.** Two dimensional diagram showing the straightforward libertarian analysis of free will, given the sentence 'some agents have free will'.

Now let's move onto the two dimensional table for the straightforward compatibilist concept of free will (see Figure 3). This two-dimensional table is simple. No matter what the actual world turns out to be like, we judge of all other worlds, regardless of whether they are deterministic or

not, that, insofar as they contain agents with compatibilist or libertarian powers, they contain free agents. Hence 'some agents have free will' comes out as true in every cell of the table.

Once again, it is important to highlight that from the index of a deterministic actual world there are no differences in free will judgments between the straightforward compatibilist concept and the conditional concept of free will (compare Figure 1). This means that differences in free will judgments between the straightforward compatibilist concept and conditional concept of free will are only found when considering people's judgments indexed to an indeterministic actual world. That's because, according to the conditional concept of free will, compatibilist powers are sufficient for free will only when determinism is true actually. Conversely, according to the straightforward compatibilist concept of free will, compatibilist powers are sufficient for free will even from the perspective of an indeterministic actual world. Keep these differences in the back of your mind for now as they will become important as we progress through this chapter. Next, I want to describe some important features of the conditional concept of free will and then provide some *prima facie* reasons for thinking that our free will concept is a conditional one.

### 2.1.2 Exploring the conditional concept of free will

So far I have described what a conditional concept is and provided my conditional analysis of our free will concept. Remember that the major task that I have set myself in this chapter is to argue that the folk concept of free will is a compatibilist concept. This is contrary to the current orthodoxy in the free will literature that thinks the folk concept is an incompatibilist one. Now I will explain what makes the conditional concept of free will a compatibilist concept. Then I will describe how the conditional concept of free will can accommodate the seeming prevalence and tenacity of people's incompatibilist intuitions about free will. Finally, I will provide some *prima facie* reasons for thinking that our concept of free will has a conditional structure.

*2.1.2.1 What makes the conditional concept of free will a compatibilist one?* The reason why the conditional concept is a compatibilist concept is because no matter how things turn out actually to be - with regard to the world being deterministic or not - to the extent that we possess this concept, we will judge that we possess free will. This is what people in two dimensional terms would call having a necessary A intension, as once you hold fixed the compatibilist powers and libertarian powers in all these worlds, all worlds considered as ways things might actually be contain agents with free will.[17] In order to get clear on this let's once look again at the two dimensional diagram for the conditional analysis of free will (see Figure 4).

[17]Using terminology developed by Jackson (1998). Chalmers (1996) uses the term 'primary intension'.

| | | Possible World: I | Possible World: D |
|---|---|---|---|
| Actual World | I | T | F |
| | D | T | T |

**'Some agents have free will'**

**Figure 4.** Two dimensional diagram showing the conditional analysis of free will, given the sentence 'some agents have free will'. The counterfactual intension (C-intension)[18] is highlighted green and the actual intension (A-intension) is highlighted red.

We can read the A intension in the two dimensional diagram by reading top-to-bottom along the left-to-right diagonal (see Figure 4). So even if determinism is actually true, and we only possess compatibilist powers, we will judge that we are free. On the other hand, if indeterminism is actually true, and we possess libertarian powers, we will judge that we are free and that indeterminism and libertarian powers are necessary for free will. So there's something we could discover, if the conditional story is correct, which would make us think that indeterminism and libertarian powers are necessary. But this doesn't mean the conditional concept of free will is not a compatibilist one, because nothing we could discover about how things are actually that would make us judge that actually there's no free will. (Remember, we're holding fixed here that there are actually either compatibilist or libertarian powers. So my claim is just that nothing we could discover about determinism would lead us to judge that we are not free, given that we have one or other of those powers. If we don't hold fixed that we have one or other of those powers, of course, then there might well be things we could discover, that world lead us to think that we are not free. Indeed, we can think of the problems raised by neuroscientists as being of exactly of this kind: as aiming to show that in fact we don't even have compatibilist powers.) So *with regard to the world being deterministic or not*, free will is compatible with anything that we could discover about how things actually are.

*2.1.2.2 What explains the prevalence and tenacity of people's incompatibilist intuitions about free will?* One important issue for anyone who thinks the folk concept of free will is compatibilist is to explain how we ever became convinced in the first place that the folk concept was incompatibilist. The conditional analysis of free will offers up an explanation of why this might be. If people think

[18]Again, using Jackson's (1998) terminology. Chalmers' (1996) refers to this as a secondary intension.

that the actual world is indeterministic and contains agents with libertarian powers, then they will judge not only that we are free, but also that any deterministic possible worlds containing only agents with compatibilist powers have no one with free will in them. In which case, to the extent people are confident that actually the world is indeterministic and there are libertarian powers, they should be expected to deny that compatibilist powers are sufficient for free will.[19] It is worthwhile to note that this is what, in two dimensional terms, is called having a contingent C intension: because from the perspective of a world where indeterminism is true, some, but not all counterfactual worlds will contain agents with free will (see Figure 4).

Of course in most of the literature the distinction between judging of the actual world that it is deterministic and that indeterminism is still a necessary condition for free will, and judging of the actual world that is indeterministic, and that this indeterminism is a necessary condition for free will, is not made.[20] What actual philosophers of free will, embedded and entrenched in their philosophical views, would judge when this distinction is drawn is not something we have tested. Nevertheless, I do think this distinction makes a difference to the judgments of ordinary agents as will become clearer as we progress.

*2.1.2.3 Why think the folk account of free will is a conditional account?* Something I find striking about the free will literature is that there are very few philosophers who take libertarianism to be a conceptual truth, and who think that the actual world is deterministic and thus are error theorists about free will. There seems to me at least to be some connection between their take on what the actual world is like, and their view about the nature of free will. Equally though, there seem to be very few compatibilists who judge that, if at some world the libertarian powers are instantiated, then those worlds do not contain free actions. So unlike the libertarian, they do not take the presence of compatibilist powers to be necessary for free will.

Some suggestive preliminary observations also come from observing the free will judgments of undergraduate philosophy students in metaphysics classes. Libertarians who become persuaded by arguments, not that compatibilism is prescriptively correct as an analysis of free will, but rather, simply that the actual world is in fact deterministic, swiftly seem to abandon their libertarianism and embrace compatibilism. With that said, they hanker after the

[19]As a descriptive matter of fact it does seem that the overwhelming majority of people think that the actual world is indeterministic. For example, Nichols and Knobe (2007) found over 90% of participants chose the vignette describing an indeterministic universe, not a deterministic universe, as being most like the actual world. Similarly, my own results (see §2.3) found 81.6% of participants selected the indeterministic universe as being most like the actual universe. I will return to these results shortly when I discuss results from experimental philosophy, including my own research.

[20]Peter Van Inwagen (1983) is the only theorist I have come across who appears to identify this distinction and thinks that the actual world is indeterministic and that this indeterminism is necessary for free will. In the very last paragraph of his book *An Essay on Free Will* he writes:

> "…it is conceivable that science will one day present us with compelling reasons for believing in determinism. Then, and only then, I think should we become compatibilists." (pg. 223)

libertarian view inasmuch as, should they start to give higher credence to the actual truth of indeterminism, they tend to shift back.[21] The observation that there are relatively few error theorists about free will and that the substantive analysis of free will tends to shift with judgments about the metaphysics of the actual world is very suggestive.

If these suggestive thoughts are correct, then there are a number of interesting upshots. One crucial upshot is one I have flagged already in this chapter. In order to empirically determine whether the folk have compatibilist or incompatibilist concepts of free will, it will be essential to determine what their views are about the actual truth of determinism, and to ensure that accidental cues in the vignettes provided to them are not determining the index at which they take themselves to be located in a way that experimenters did not intend, or did not recognise. In the next section I will show that this has three important consequences for the interpretation of existing work on the folk concept of free will. The first is that apparently differing judgments about whether the folk are compatibilists or incompatibilists may be artefacts of features of the experimental set up that have not hitherto been regarded as important. The second is that in fact all this data is consistent with a compatibilist reading of the concept of free will insofar as it is a conditional compatibilist account. The last is that the methodology so far deployed is unable to fully distinguish between conditional and straightforward accounts.

### 2.1.3. Conclusion: What is a conditional account of our concept of free will?

In summary, in this chapter I have set about the task of arguing that the folk concept of free will is compatibilist, albeit of a conditional kind. A conditional concept is a single concept that has an indexical component which fixes what is necessary for something depending on what the actual world offers up. Appling this thought to the case of free will I have proposed the following conditional analysis of free will:

> If the actual world is indeterministic, and agents have libertarian powers, then these libertarian powers are what free will is and must be. Else, if the actual world is deterministic, and agents have their preferred compatibilist powers, then compatibilist powers are what free will is.

The conditional concept is a compatibilist concept, in some sense, because no matter how things turn out actually to be with respect to the world being deterministic or not, to the extent that we possess this concept, we will judge that we possess free will. Further, the conditional concept can accommodate the prevalence and tenacity of people's incompatibilist intuitions, intuitions whose

[21]This observation come from an ongoing longitudinal study that I am currently conducting with David Braddon-Mitchell that tracks undergraduate free will judgements and belief about whether determinism or indeterminism is actual over the course of the teaching semester.

presence resulted in the orthodox view that the folk concept of free will is incompatibilist. Insofar as people judge that the actual world is indeterministic and contains agents with libertarian powers, then not only will they judge that we are free, but also that any counterfactual deterministic possible worlds containing only agents with deterministic powers will not contain free will.

Some observations of free will theorists and undergraduate students suggests that there might be a connection between what people think the actual world is like, and their view about the nature of free will. This has important consequences for the project of determining what our free will concept is like. From the index of the indeterministic actual world there are no differences in free will judgments between the straightforward libertarian concept and conditional concept of free will. Further, from the index of the deterministic actual world there are no differences in free will judgments between the straightforward compatibilist concept and conditional concept of free will. This means that it is essential to track people's beliefs about the actual truth of determinism in order to know whether they have an incompatibilist concept of free will, or a compatibilist concept of free will of a straightforward or conditional kind.

### 2.2. Experimental philosophy and the conditional concept of free will.

Previously, I suggested that there is excellent evidence from experimental philosophy that the folk concept of free will is both a compatibilist concept (e.g. Nahmias et al., 2005; 2006) and an incompatibilist concept (e.g., Nichols and Knobe, 2007). In this subsection I will do three things. First, I will to explain why current results regarding our concept of free will in the experimental philosophy literature only superficially differ. Roughly, that's because I think the differences in people's responses in these studies are due to subtle methodological differences. Second, I will explain that these subtle methodological differences are correctly identified as being important with respect to people's concept of free will once it has been identified that our concept of free will might be a conditional concept. Thus, I will explain why existing experimental results are consistent with our underlying concept of free will being a compatibilist concept albeit a conditional one. Finally, I will go on to show that current results regarding our concept of free will in the experimental philosophy literature can only go part of the way to determining whether the concept is a conditional concept or not (e.g., Roskies and Nichols, 2008). From this I will articulate what a strong data signal for conditionality in the folk concept of free will would look like. This hypothesized data signal for the conditional concept will serve as the basis of the next section (§2.3) where I will present evidence of this strong data signal from my own investigation.

If my results are right, then the folk concept of free will is a conditional concept and thus the folk are free will compatibilists.

**2.2.1 Experimental philosophy and the folk concept of free will.**

I will start by simply describing two sets of conflicting results. One set of results appears to support the idea that the folk conception of free will is compatibilist. The other appears to support the idea that the folk conception is incompatibilist. *Prima facie*, then, you might think that either one or both of the studies is badly flawed, or they have tested, by chance, very different populations. I will argue that neither of these is true. Both studies are fine as far as they go, but if we examine them closely we find subtle differences in the methodological set up that could explain the differences in the results. This difference in the results means that both studies are compatible with the underlying concept in the community being the conditional one. In addition, I will argue that the studies themselves, though they are not set up in a way which can decisively detect the conditional concept, nevertheless might provide weak evidence that it is indeed the conditional concept at work.

*2.2.1.1. The folk concept of free will is compatibilist.* Perhaps the best evidence showing that the majority of the general population are compatibilists comes from seminal papers by Nahmias, Mossis, Nadelhoofer, and Turner (2005; 2006).[22] Consider the following vignette of an agent's action embedded within a description of our world being deterministic:

> "Imagine that in the next century we discover all the laws of nature, and we build a supercomputer which can deduce from these laws of nature and from the current state of everything in the world exactly what will be happening in the world at any future time. It can look at everything about the way the world is and predict everything about how it will be with 100% accuracy. Suppose that such a supercomputer existed, and it looks at the state of the universe at a certain time on March 25th, 2150 A.D., twenty years before Jeremy Hall is born. The computer then deduces from this information and the laws of nature that Jeremy will definitely rob Fidelity Bank at 6:00 PM on January 26th, 2195. As always, the super computer's prediction is correct; Jeremy robs Fidelity Bank at 6:00 PM on January 26th, 2195."

When college students were presented with such a vignette, 76% of college-aged participants judged that the agent in the scenario acted using his own free will. Given some of the technical language involved in the description of determinism (i.e., laws of nature) the experiment was repeated using varying descriptions, either by describing the universe as re-started with the same initial conditions or by appealing to genetic and environmental influences. Each time, though,

[22]See also Murray & Nahmias (2014); Nahmias, Coates, & Kvaran (2007); and, Nahmias & Murray (2010).

the majority of the participants continued to respond that the agent in each scenario was acting using his own free will. Most people, then, appear to straightforwardly respond that an agent acts freely in an actual deterministic world and so it seems most people must be compatibilists.

*2.2.1.2. The folk concept of free will is incompatibilist.* Now let's move onto what I think is some of the best evidence that the general population are incompatibilists, which comes from the seminal paper by Nichols and Knobe (2007).[23] They proposed a performance error model to explain Nahmias and colleagues results, according to which our actual concept of free will is incompatibilist, but we begin performing in error and responding as compatibilists when considering real life vignettes in which participants are invited to consider that determinism is true. To test this, college-aged participants were provided with descriptions of two hypothetical universes, Universe A, a deterministic universe in which everything including human decision making is completely caused by events tracking all the way back to the beginning of the universe, and Universe B, an indeterministic universe where everything with the exception of human decision making is completely caused by past events. When asked which universe was most like the actual universe over 90% responded that Universe B was the most like our universe: most people believe that the actual universe is indeterministic. Then when asked whether participants in Universe A were morally responsible (a proxy for free will), 86% of participants responded no, the incompatibilist response. Most people, then, straightforwardly respond that there is no free will in a counterfactual deterministic world and so most people must be incompatibilists.

However, this study went further; people's responses drastically changed when they were given an emotionally valenced real life scenario embedded in Universe A. Consider the following vignette:

> "In Universe A, a man named Bill has become attracted to his secretary, and he decides that the only way to be with her is to kill his wife and 3 children. He knows that it is impossible to escape from his house in the event of a fire. Before he leaves on a business trip, he sets up a device in his basement that burns down the house and kills his family."

In this case 72% of participants gave a compatibilist response, suggesting that the agent was fully morally responsible. Nichols and Knobe put this down to a performance error on the part of the participants, suggesting that they were being psychologically swayed by the emotional valence of the scenario that was presented. While there is no doubt that emotional valence can impact our judgments, it cannot be the case that all compatibilist responses are the result of a performance error triggered by emotional processing. After all, Nahmias and colleagues also found a majority of participants reporting compatibilist intuitions in concrete but emotionally neutral vignettes

[23]See also Rose & Nichols (2013) and Nichols (2004).

(i.e., jogging). Instead, it seems that most people, when asked straightforwardly, respond that there is free will in real-life cases regardless of emotional valence. I will return to discuss the significance of this result shortly.

*2.2.1.2. Explaining away the apparently conflicting results in experimental philosophy.* Let's consider three major kinds of differences that we might point to in order to explain these differences in results. Once I have done this I will explain the significance of each of these factors in explaining what is going on.

First there might be some systematic differences in the sampling populations that are being asked to give judgments in the various conditions. Perhaps Nahmias and colleagues just happened to sample, by chance, an in some repeated fashion, those people in the general population that possess a compatibilist concept of free will, and on the other hand, perhaps Nichols and Knobe have just happened to sample those people in the general population that possess an incompatibilist concept of free will.

Second, perhaps one or the other samples of people in these studies are making some performance error of some kind. For example, perhaps Nichols and Knobe (2007; though see also Rose and Nichols, 2013) are right and people are making a performance error because of the emotional valence or the real-world setting. Or perhaps, the opposite is true, as Nahmias and Murray (2010; though see also Murray and Nahmias, 2014) have suggested, and people only give incompatibilist judgments because they don't correctly understand the implications of determinism for their abilities to cause the actions that they do.

Finally, perhaps it might be explained by the subtle difference in the way that people are questioned about whether agents have free will or not. So for example, the people in Nahmias' and colleagues studies (as well as in Nichols' and Knobe's for that matter) respond that people in real-world deterministic scenario are free and that, conversely, people that are located in some other deterministic scenario are not free.

For my purposes what is interesting is that after having identified the possibility that our concept of free will might be a conditional concept, there is really no reason to suspect that different research groups by chance have come across different populations of people who have different concepts of free will, or that some participants are responding in error. In fact, the observed pattern of responses is exactly what you would expect if people possess a conditional concept of free will. Recall from the previous section what the conditional concept of free will says. If the actual world is indeterministic, and agents have libertarian powers, then these libertarian powers are what free will is and must be. Counterfactually we would, in this condition, judge that all and only indeterministic worlds containing agents with libertarian powers contain

agents with free will. On the other hand, if the actual world is deterministic, and agents have their preferred compatibilist powers, then compatibilist powers are what free will is, and counterfactually we will judge compatibilist powers to be sufficient for free will. We can explain the current patterning of results perfectly with this conditional analysis of free will at hand. Let's go back through the results again.

First, why do Nahmias and colleagues find that the most people responding to vignettes situated in the actual deterministic world are free? Because given that we have a conditional concept of free will, if the actual world is deterministic then compatibilist powers are sufficient for free will. Conversely, why do Nichols and Knobe find that the majority of people responding judge that agents in some hypothetical deterministic vignette are unfree? Recall that their preliminary results showed that the vast majority of people think the actual world is indeterministic. Given that we have a conditional concept of free will, if the actual world is indeterministic, then counterfactually we only judge that indeterministic worlds contain agents with free will. Thus, as the overwhelming majority of people in their study think the actual world is indeterministic, if their concept is conditional they will judge that agents in counterfactual deterministic worlds are unfree. In both cases, what we wold predict if our concept is conditional is exactly what we find in these experiments.

There might even be a very weak data signal from Nichols and Knobe (2007) that the conditional concept is at work in these results. Why do Nichols and Knobe also find that the majority of people responding judge that agents in some real-life deterministic scenario are free? Because if people have a conditional concept of free will we will predict that if they think that the real-life scenario is occurring in the deterministic actual world, then they will judge that agents in the actual deterministic world are free.

This is not to say that people's judgments are not sensitive to factors such as emotional valence. If that's the only factor that people are sensitive to then these results are no evidence for the conditional concept. Still, as we will discuss shortly, Roskies and Nichols (2008) do find people's free will judgments are sensitive to the distinction between actual and counterfactual scenarios, so these results are consistent with people responding from a conditional concept of free will.

Of course, it remains open that the people responding in each of these studies just *happen* to consistently respond as though they are free will compatibilists or as though they are free will incompatibilists. However, if that were true it would be a remarkable fluke. Instead, it seems to me far more likely that subtle differences in the probe questions that are being used to illicit

people's judgments are causing people to respond from different indices, whilst deploying a conditional concept.

Further, it remains open that perhaps one group of people in one or another of these studies is making some performance error of some kind. But why think that? Of course it is possible that people responding to the deterministic cases are making an error, just as it is possible that people in the indeterministic cases are making an error. But then the question turns into the question of which sets of responses we ought to think actually issue from people's underlying concept of free will. It seems much better, and more plausible by my lights, if the data signal from both sets of studies is right, and what's more both issue from the same underlying conditional concept of free will. Then people are sincerely responding in both cases, and that means that people's judgments are sensitive not only to the vignette they are being asked to consider, but also to the way they are meant to consider the world as being actually.

Up to this point the following things should be clear. There is a clear signal of people responding as though they possess a compatibilist concept of free will and as though they possess an incompatibilist concept of free will. *Prima facie* this suggests that the experimental results from experimental philosophy are inconsistent; but this inconsistency is merely surface level. Closer examination of the studies reveals subtle differences in the methodological set up. People straightforwardly respond as though they have a compatibilist concept of free will when they are being asked to consider the vignette in a real-world deterministic scenario, and straightforwardly as though they have an incompatibilist concept of free will when they are being asked to consider the vignette in a hypothetical (i.e., counterfactual) deterministic scenario. Both these results are consistent with, and predicted by, the thought that people have a conditional concept of free will.

**2.2.2 Bringing free will back down to earth.**

Roskies and Nichols (2008; though see also Björnsson, 2014) also noticed a slight, yet significant, difference in the experimental materials used in the experimental philosophy literature that could have generated conflicting findings. While Nahmias and colleagues situated some of their determinism scenarios within the actual world, Nichols and colleagues situated them in hypothetical worlds. In order to confirm their suspicion that participant's free will judgements under determinism differed as a result of where they were being evaluated, participants were evenly split between considering the deterministic scenario as the actual world or as some other hypothetical world. Consistent with the author's hypotheses, where the deterministic scenario was situated significantly impacted participant's free will judgements. Participants who made free will judgements when the deterministic world being evaluated was our own were significantly

higher relative to participants who made free will judgments when the deterministic world being evaluated was not our own.

Roskies and Nichols, following Braddon-Mitchell (2003), in passing suggest that these results might suggest that the folk concept of free will takes a conditional form. As described earlier, it should be easy to see why. When a vignette is taken to describe the actual world, we should expect, if people are deploying a conditional concept, that they will judge that agents act freely in the deterministic world considered as actual, and that people will judge that agents act unfreely in the counterfactual deterministic world. The reason they are inclined to judge people in the counterfactual deterministic world as being unfree is because people as a matter of fact believe that the actual world is indeterministic and so think, unless told otherwise, that indeterminism is a necessary condition for free will. So far so good; but this evidence is only consistent with the folk having a conditional concept of free will. The reason these results are only consistent with the conditional concept and do not show that people in fact possess a conditional concept of free will is because we do not have data and responses to all the conditions necessary to determine whether or not there is a conditional concept.

*2.2.2.1 Why are people unsure about free will?* Let's assume for the moment that the majority of people in this study think the actual world is indeterministic. If these same people also have a conditional concept of free will then when evaluating the deterministic vignette as actual they should straightforwardly respond that there is free will. On the other hand, when evaluating the deterministic vignette as a hypothetical they should straightforwardly respond there is no free will. It is important to note that Roskies' and Nichols' conditionality suggestion leans on the fact that people's free will judgments for the deterministic actual world are higher than people's free will judgments for the hypothetical deterministic world. However, as a descriptive matter of fact, the mean free will score for a deterministic actual world was 4.3 on a 7-point Likert scale. A score of 1 would indicate complete agreement that there is free will in that world, and a score of 7 would indicate completely disagreement there is free will in that world. *Prima facie*, then, people aren't straightforwardly responding as though there is free will in the deterministic actual world. Instead, while they straightforwardly judge that there is no free will in the hypothetical deterministic world, they are unsure about whether there is free will in the deterministic actual world.

Here are two hypotheses about what might be going on. Unfortunately, I cannot get traction on exactly which hypotheses is most likely, in the Roskies and Nichols study, as there is no indication of how people's responses are distributed across the 7-point Likert scale. One way to obtain middling mean judgements on a Likert scale is by people being evenly distributed

between thinking that there is, and is not, free will in the deterministic actual world condition. If that's right, then these results might not be consistent with the folk concept of free will being a conditional one. That's because you could generate their significant result with only a significant minority of people possessing the conditional concept of free will, and the majority of people possessing either the straightforward compatibilist or straightforward incompatibilist concept of free will. Provided there are at least some people who possess the conditional concept of free will, then the different judgments they offer up in the deterministic worlds taken as actual and counterfactual will be enough to generate a mean score of 4.3 (a score that is further towards taking the deterministic world to contain free will, than taking it not to).

Secondly, perhaps people's responses in the determinism as actual condition - the condition in which they take the deterministic world to be actual - are clustered around 4 because people are in fact genuinely unsure, in some sense, whether or not there is free will in that condition. As a matter of fact, I think this is the most likely hypothesis as it's what I find in my own investigations, which I will describe in §2.3. I don't think this result should be surprising, especially if what underlies people's judgments is a conditional concept of free will. Here's why. Imagine someone who genuinely believes the actual world is indeterministic and is asked to evaluate whether there is free will in an actual deterministic world. It's extremely unlikely that in order to perform the evaluation they change their belief so that they now believe that the actual world is deterministic. Instead, what they will likely do is simulate in their mind how they would respond given a counterfactual belief that the actual world is deterministic. Importantly, this simulating does not mask what people actually believe, which is what explains why people are genuinely unsure how to respond. If people possess the conditional concept and believe the actual world is indeterministic then they should think that indeterminism and libertarian powers are necessary for free will. So according to their actual beliefs there is no free will in the actual deterministic world. But, if they possess the conditional concept and correctly simulate how they would respond given the counterfactual belief that the actual world is deterministic, then they should also think the compatibilist powers are sufficient for free will. So according to their simulated counterfactual belief there is free will in the actual deterministic world. Thus there is a conflict in response caused by one response generated in accordance with the actual belief, and one response generated in accordance with the simulated counterfactual belief.

One thing that this suggestion highlights is the future need for longitudinal studies that track people's free will judgements over time and across changes in belief about what the actual world is like. That's because if I am right then when people's beliefs about what the actual world is like change, they should no longer be unsure about how they should respond in certain

conditions. For example, imagine once again someone who believes indeterminism is true actually and is unsure whether or not there is free will in a deterministic actual world. If they possess a conditional concept of free will and come to believe that the actual world is deterministic, then at that point they should straightforwardly think that there is free will in the actual deterministic world. From pilot data that we have tracking undergraduate philosophy students in metaphysics classes this appears to be happening. However, these investigations are in the earliest stages and ongoing so I will not lean on them further in this thesis.

*2.2.2.2. The missing response link to detecting conditionality.* I will call the prediction make by the conditional concept of free will that people who think the actual world is indeterministic will come to think that there is free will in an actual deterministic world the **weak signal for conditionality**. That's because I think there is another plausible route to this same result. Specifically, someone with a straightforward incompatibilist concept of free will, upon learning that the actual world is deterministic, might simply revise their concept of free will so that they now have a straightforward compatibilist concept. Importantly, the conditional concept account that I have been advocating in this chapter makes an incredibly novel prediction that sets it apart from this revisionary explanation. According to the revisionary story, if someone then switches to believing that the actual world is indeterministic again, there is no guarantee they will re-revise their concept. However, the conditional account predicts that if someone believes the actual world is deterministic, then when they come to believe the actual world is indeterministic then they will begin responding like an incompatibilist. While this might also be explained by someone switching their concept back, I think this is a stronger signal as it predicts that people's judgements will change not for the world considered as actual, but for counterfactual worlds from the perspective of the world now believed to be actual.

Thus, it seems to be that the strongest signal of the conditional concept comes not from people who believe that the actual world is indeterministic, but rather those who believe the actual world is deterministic. These are people who when asked to consider the actual deterministic world and other counterfactual deterministic worlds will respond that there is free will in each of these cases. Importantly, when they are asked to consider the actual world as being indeterministic they will judge that agents in the indeterministic world considered actual would are free. Critically, though, from the point of view of the actual world considered as indeterministic, they will judge that agents in counterfactual deterministic worlds are *unfree*. I will call this prediction the **strong signal for conditionality**. In the next section I will present evidence for both the weak and strong signal for conditionality from my own investigation.

Of course, if my explanation of the middling result (i.e., the 4.3 mean response to the deterministic actual world) found in Roskies and Nichols (2008) is right, then people who genuinely believe the actual world is deterministic, when asked to imagine that the actual world is indeterministic, might not straightforwardly judge that there is no free will in a counterfactual deterministic world. Instead, they might respond in a manner that is consistent with them being unsure. That's because given their actual belief that the world is deterministic, compatibilist powers are sufficient for free will and so there is free will in any deterministic world that contains compatibilist powers. But according to their simulated counterfactual belief that the actual world is indeterministic, there is no free will in any deterministic world. That's because, according to the conditional concept of free will if the actual world is indeterministic then indeterminism and libertarian powers are necessary for free will.

While the result of people being unsure is an overall weaker result than if people straightforwardly respond according to the conditional concept, provided people are only unsure in the conditions described by the weak and strong signal for conditionality then I still think this is relatively strong evidence of conditionality. What it does do is further motivate the need for future longitudinal studies tracking people's free will judgments over time and across changing beliefs about what the actual world is like.

**2.2.3 Conclusion: Experimental philosophy and the folk concept of free will.**

In summary, in this section I have shown that there is a clear signal in experimental philosophy of people responding as though they possess a compatibilist concept of free will and as though they possess an incompatibilist concept of free will. However, any differences in results that purport to show the folk are compatibilists and those that purport to show that the folk are incompatibilists are superficial and due to differences in methodological set up. These differences in methodological set up are correctly seen as being important when it has been identified that the folk concept of free will might be a conditional one. People should be expected to straightforwardly respond as though they have a compatibilist concept of free will when they are being asked to consider a deterministic actual world. Further, they should be expected to straightforwardly respond as though they are an incompatibilist when they are being asked to consider counterfactual deterministic worlds. That's because most people believe that the actual world is indeterministic and according to the conditional concept of free will, from the perspective of an indeterministic actual world no deterministic world has free will.

Roskies and Nichols (2008) identified a weak signal for conditionality. In cases where people judge there was no free will that judgment was only certain when the deterministic world was considered as counterfactual and not as actual. Unfortunately, we do not have responses to

all the conditions necessary to determine whether or not there is a conditional concept. Further, we are missing some key measures of statistical variance that would allow us to at least go some way to ruling out competing hypotheses. Specifically, the mean free will response by participants to the deterministic actual world was 4.3, which is a middling response on a 7-point Likert scale. While I think this is consistent with an underlying conditional concept, with people unsure due to competing dispositions to respond according to what they actually believe versus what their response would be if the actually believed different, it is also consistent with people being evenly distributed between thinking that there is free will or not, in those conditions.

The strongest data signal for conditionality comes not from people who believe that the actual world is indeterministic, but rather, those who believe the actual world is deterministic. Critically, from the point of view of the actual indeterministic world people who believe that the actual world is deterministic should be expected to judge that agents in the counterfactual deterministic worlds are unfree. As a matter of fact, I find evidence of both the weak and strong signal for conditionality, which I will present next. If I am right then this is strong evidence that the folk concept of free will is a conditional one, in which case the folk are free will compatibilists.

### 2.3. Detecting conditionality in the folk concept of free will[24]

According to the conditional concept of free will what people take to be necessary for free will is indexed to what they believe the actual world is like. This feature makes two key predictions regarding people's free will judgments to various conditions, which I have called the weak and strong signals for conditionality.

---

[24]You might query why an experimental philosophy investigation is necessary at all. After all, Dunaway, Edmonds and Manley (2013) found that professional philosophers are remarkably good at predicting the results of experimental philosophy studies on examining folk judgements for a variety of topics, including: causation, moral responsibility and intentionality. Depending on the exact experimental study examined, the proportion of professional philosophers who could correctly predict the study's results was between 77% and 95.8%. On the basis of this result Dunaway et al. offer a tentative conclusion. If the results of x-phi studies can be accurately predicted by professional philosophers, then when considered with other factors such as resource scarcity, professional philosophers may have ample reason not to conduct the relevant experimental philosophy research. Put another way, if the results of x-phi research assessing folk judgments are unsurprising and predictable then such research might be superfluous.

I'm less confident. There are two possible questions that Dunaway could have addressed using the data set they collections. First, how good *collectively* are philosophers at predicting folk judgments? Second, how good are *individual* philosophers at predicting folk judgments. The original analysis performed by Dunaway et al. only answered the first of these questions, and showed that professional philosophers as a collective are good at predicting folk judgements. A reanalysis performed by Liao (2016) answers the second question, and found that roughly half (51.9%) of all professional philosophers in Dunaway et al. got at least one prediction wrong. Some readers might point out that the fact that half of all professional philosophers are getting all their predictions right is still impressive, and they are right. However, we have no a priori way of knowing which professional philosophers will get which judgements right and wrong and for which cases. The only way to know this is by performing the relevant experimental philosophy.

The weak signal for conditionality is what Roskies and Nichols (2008) identified might be present in their data (and might also be present in Nichols & Knobe, 2007). If people who believe the actual world is indeterministic possess a conditional concept, and are asked to evaluate the actual deterministic world, they should be expected to respond that there is free will in such a world. That's because according to the conditional concept, indeterminism and libertarian powers are only necessary for free will if they obtain actually. This weak signal for conditionality distinguishes the conditional concept from the straightforward libertarian concept that predicts people will respond that there is no free will in the actual deterministic world because indeterminism and libertarian powers are necessary for free will irrespective of what the actual world is like.

On the other hand, the strong signal for conditionality makes the following prediction. If people who believe the actual world is deterministic possess a conditional concept, then when they are asked to evaluate a counterfactual deterministic world from the perspective of an indeterministic actual world, they should be expected to respond that there is no free will in that world. That's because according to the conditional concept, indeterminism and libertarian powers are necessary for free will if the actual world is indeterministic. The strong signal for conditionality distinguishes the conditional concept from the straightforward compatibilist concept because according to the straightforward compatibilist concept we should expect people to respond that there is free will in the counterfactual deterministic world because compatibilist powers are sufficient for free will irrespective of what the actual world is like.

Now I will present what I think is good evidence of both the weak and strong signals of conditionality in free will judgments. Unfortunately, most people do not straightforwardly respond in the manner predicted if they were deploying a conditional concept of free will. Instead, they appear unsure of how to respond, in a manner that is consistent with what was observed by Roskies and Nichols (2008). While this is weaker evidence than I would have liked, I will argue that it still gives us a strong reason to think that the folk concept of free will is a conditional one. That in turn also gives us a strong reason to think the folk are free will compatibilists.

**2.3.1. Empirical evidence of conditionality.**

I will report the results of my investigation in two parts. Here I will present just the evidence for the weak and strong signals of conditionality. For those that are interested I will present the methods and materials behind these results in §2.3.2. I will start by presenting evidence of the weak signal for conditionality in people who believe the actual work is indeterministic. Then, I will present evidence of the strong signal for conditionality in people who believe the actual

world is deterministic. After that I will interpret what these results mean for the conditional concept account.

*2.3.1.1 The weak signal for conditionality.* Figure 5 below presents the mean free will responses of people who believe the actual world is indeterministic. A value of 7 indicates that people completely agree that there is free will in the world being evaluated, and a value of 1 indicates that they completely disagree that there is free will in the world being evaluated. People who *in fact* believe the actual world is indeterministic, straightforwardly respond that there is free will in an indeterministic actual world ($M$ = 6.37, $SD$ = 1.19) and in a counterfactual indeterministic world, from the perspective of a deterministic actual world ($M$ = 6.01, $SD$ = 1.63). Further, they straightforwardly respond there is no free will in a counterfactual deterministic world ($M$ = 2.03, $SD$ = 1.70). Of course, these judgments are all predicted by the straightforward libertarian concept of free will as well. According to the weak signal of conditionality we should expect people to respond that there is free will in a deterministic actual world. However, consistent with Roskies and Nichols (2008) I found that people who in fact believe the actual world is indeterministic are unsure whether or not there is free will in this condition ($M$ = 3.98, $SD$ = 1.02). While there are some individuals who judge otherwise, most people's judgements are clustered around 4. I will return to discuss this finding shortly.

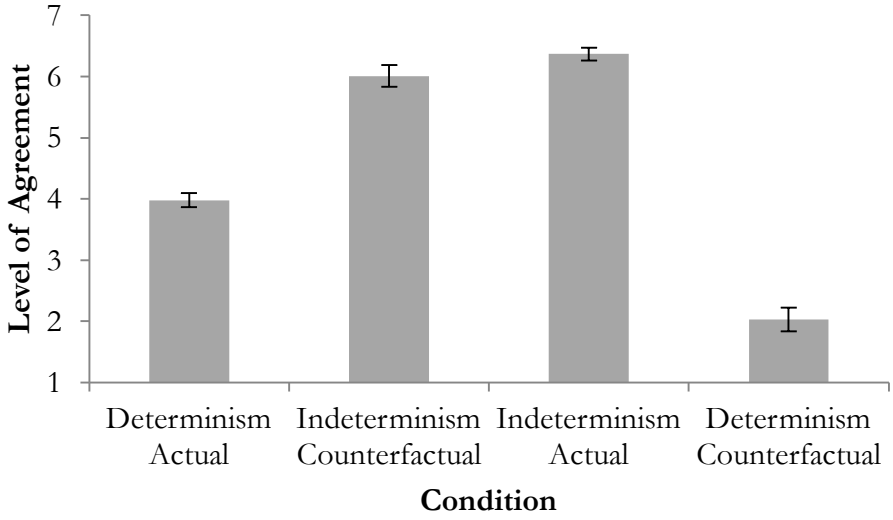


**Figure 5.** Levels of agreement with the statement "If these scientists are right, then the people in Universe A [/B] have free will" for those participants who believe the actual universe is indeterministic. Errors bars represent standard error.

*2.3.1.2 The strong signal for conditionality.* Figure 6 below presents the mean free will responses of people who in fact believe the actual world is deterministic. People who believe the actual world is deterministic straightforwardly respond that there is free will in a deterministic actual world ($M$ = 5.94, $SD$ = 1.06) and an indeterministic actual world ($M$ = 5.91, $SD$ = 1.18). In addition, they straightforward respond that there is free will in a counterfactual indeterministic world, from the perspective of a deterministic actual world ($M$ = 5.67, $SD$ = 1.14). Of course these judgments are predicted by the straightforward compatibilist concept of free will as well. According to the strong signal of conditionality, if people deploy the conditional concept we should expect them to respond that there is free will in a counterfactual deterministic world, from the perspective of an indeterministic actual world. While people don't straightforwardly judge that there is no free will in this condition, as with the weak conditional signal, people are unsure whether or not there is free will in this condition ($M$ = 4.06, $SD$ = 1.26). While some individuals respond otherwise, most people's judgements are clustered around 4.[25]

[25]I think that Figure 5 and Figure 6 with standard errors represented are sufficient to represent evidence of the weak and strong signal of conditionality. However, some readers might be interested whether the mean response for each condition significantly different between those who think the actual world is indeterministic and those who think the actual world is deterministic. Due to the severe discrepancy in numbers between those who think the actual world is indeterministic and deterministic the decision was made to compare conditions between groups using a series of between-groups t-tests with a Bonferroni adjusted alpha level. Significant differences were only observed in those conditions associated with the weak and strong signal for conditionality. That is, there was a significant difference in the determinism actual case $t(96)$ = 7.3.64, $p < .001$. People who believe the actual world is deterministically straightforwardly responded there was free will, people who believe the actual world is indeterministic responded unsure. And, there was a significant difference in the counterfactual determinism case counterfactual determinism case $t(96)$ = 4.777, $p < .001$. People who believe the actual world is indeterministic straightforwardly responded there was no free will, people who believed the actual world is deterministic responded unsure. There was no significant difference between people who believe the actual world is indeterministic and deterministic in the indeterminism actual case $t(96)$ = 1.784, $p$ = .078, and counterfactual indeterminism case $t(96)$ = 1.456, $p$ = .396. Regardless of what people believe the actual world is like, people in these cases straightforwardly respond there is free will.

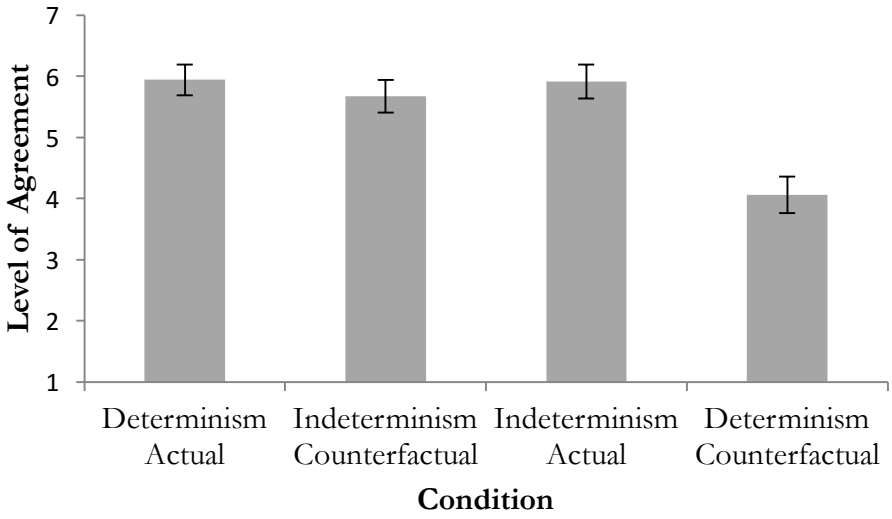


**Figure 6.** Levels of agreement with the statement "If these scientists are right, then the people in Universe A [/B] have free will" for those participants who believe the actual universe is deterministic. Error bars represent standard error.

*2.3.1.3 Interpreting the conditionality data signal.* In accordance with the conditional concept of free will, the weak signal of conditionality predicts that people who believe the actual world is indeterministic will respond that there is free will in the actual deterministic world. On the other hand, the strong signal of conditionality predicts that people who believe the actual world is deterministic will respond that there is no free will in a counterfactual deterministic world. While I didn't observe people straightforwardly responding in the manner predicted by the conditional concept, they do respond in a manner that supports the idea that we possess a conditional concept of free will. That's because I don't think it is mere coincidence that people who believe the actual world is indeterministic are unsure how to respond to an actual deterministic world. Nor do I think it's coincidence that people who believe the actual world is deterministic world are unsure how to respond to a counterfactual world from the perspective of an actual indeterministic world. Both these conditions are correctly identified as being important with respect to people's concept of free will once it has been identified that our concept of free will might be a conditional concept.

Earlier, I gave an account of why I think people might find themselves unsure about how to respond in conditions associated with the weak and strong signal for conditionality. Let's start with the weak signal for conditionality. Imagine someone who believes the actual world is indeterministic and is then asked to evaluate whether there is free will in the actual deterministic world. It's extremely unlikely that people change their beliefs about the actual world in order to perform such evaluations. Instead, what people most likely do is simulate how they would

respond if they counterfactually believed the actual world is deterministic. Importantly, this cognitive process does not mask the effects of what people actually believe, which is what explains why people are unsure about how to respond. If someone has a conditional concept and believes the actual world is indeterministic then they should also think that indeterminism and libertarian powers are necessary for free will. So according to their actual belief there is no free will in the deterministic actual world. But if they succeed in simulating what they would think if they counterfactually believed the actual world is deterministic, then they should also think compatibilist powers are sufficient for free will. So according to their simulated counterfactual belief there is free will in the actual deterministic world. Thus, there is a response conflict between free will responses generated in accordance with their actual belief, and their simulated counterfactual belief.

The very same explanation works for why we observe that people who believe the actual world is deterministic are unsure how to respond in the condition associated with the strong signal for conditionality. Imagine now someone who believes the actual world is deterministic and is asked to evaluate whether there is free will in a counterfactual deterministic world from the perspective of an indeterministic actual world. If that person has a conditional concept, then they should also think that compatibilist powers are sufficient for free will. So according to their actual belief there is free will in the counterfactual deterministic world. But if they succeed in simulating what they would think if they counterfactually believed the actual world is indeterministic, then they should no longer think that compatibilist powers are sufficient for free will. Instead they should think that indeterminism and libertarian powers are necessary for free will. So according to their simulated counterfactual belief there is no free will in the counterfactual deterministic world. As a result, there is once again a response conflict between free will responses that are generated in accordance with someone's actual and simulated counterfactual beliefs.

While the current results are weaker evidence for conditionality than I would have liked, I still think they give us a reason to think that what underlies people's responses is a conditional concept of free will. Looking forward there is a need to run confirmatory longitudinal studies. My explanation of my data is that we are relying on agents simulating what they *would* judge if they came to change their beliefs. This allows for two effects: poor simulation, on the one hand, and a conflict between outputs of simulation and straightforward judgments given how they actually take things to be, on the other hand. In a longitudinal study I will have access to agents that have actually changed their minds, so we will not be relying on their imperfect simulations of what they would do if they changed their minds, and also there will be no conflict with

judgments based on how they take things to be actually. People who believe the actual world is indeterministic and so are unsure how to respond in the actual deterministic world condition should straightforwardly respond that there is free will if they come to genuinely belief the actual world is deterministic. Further, people who believe the actual world is deterministic and then come to believe the actual world is indeterministic should be expected to straightforwardly judge there is no free will in the counterfactual deterministic world from the perspective of the indeterministic world. As I alluded to earlier, these investigations have already begun and my collaborator and I have some evidence people are no longer unsure how to respond when they genuinely change their beliefs about what the actual world is like. However, these investigations are ongoing and we need many more cases of people genuinely changing their beliefs about what the actual world is before we can be confident in these results. Nevertheless, this suggestive evidence is enticing as it would constitute more definitive proof that the folk concept of free will is a conditional one and thus definitively show that the folk are free will compatibilists.

**2.3.2. Behind the empirical result: Methods of the experimental task.**

For those interested readers, here I will describe the methods and materials I used to obtain the results described in §2.3.2. 102 people participated in the study. Participants were U.S. residents, recruited and tested online using Amazon Mechanical Turk. Four participants had to be excluded for failing to follow task instructions. This means that they failed to answer all the questions, or failed an attentional check question. The remaining sample was composed of 98 participants (aged 23-69; 44 female). Mean age 37.12 (*SD* = 10.39). Ethics approval for this study was obtained from the University of Sydney Human Research Ethics Committee. Informed consent was obtained to all participants prior to testing. The study was conducted online using Qualtrics.

At the beginning of the experimental, each participant was presented with two vignettes. One vignette described a deterministic universe the other vignette described an indeterministic universe. Participants had continued access to the vignettes throughout the experiment. They read as follows:

> Imagine a universe (Universe A) in which everything that happens is completely caused by whatever happened before it. For example, one day Katie decided she wanted to have a cup of coffee with her breakfast. Like everything else, this decision was completely caused by what ever happened before it. So if everything in this universe was exactly the same up until Katie made her decision, then it *had to happen* that Katie would decide to have a cup of coffee. Every decision Katie makes is completely caused by whatever happened before the decision – given the past, every decision *has to happen* exactly the way that it does.

> Imagine a universe (Universe B) in which everything that happens is not completely caused by whatever happened before it. For example, one day Katie decided she wanted to have a cup of coffee with her breakfast. Like everything else, this decision is not completely caused by what happened before it. So if everything in the universe was exactly the same up until Katie made her decision, it *did not have happen* that Katie would decide to have a cup of coffee. Every decision Katie makes is not completely caused by whatever happened before the decision – given the past, every decision *does not have to happen* the way that it does.

Participants were then asked: Which of these universes do you think is most like our own? Consistent with the results of Nichols and Knobe (2007), the overwhelming majority of people picked Universe B (81.6%), the indeterministic universe, and not Universe A (18.4%), the deterministic universe.

Each participant was then presented with four different conditions in random order. Each condition corresponds to one of the boxes in the two dimensional table. Participants were asked to evaluate whether there was free will in: (1) a deterministic actual world, (2) an indeterministic actual world, (3) counterfactual deterministic world from the perspective of an indeterministic actual world, and (4) counterfactual indeterministic world from the perspective of a deterministic actual world.

Here's how the setup in each condition worked. In condition (1) a deterministic actual world, participants were asked to: Imagine scientists discover that the actual universe (the one where you or I live) is exactly like Universe A. Similarly, in condition (2) an indeterministic actual world, participants were asked to: Imagine scientists discover that the actual universe (the one where you or I live) is exactly like Universe B. Meanwhile, in condition (3) a counterfactual deterministic world, participants were asked to: Imagine scientists discover that the actual universe (the one where you and I live) is exactly like Universe B. Now imagine scientists are suddenly able to observe a parallel universe. The parallel universe contains everything that the actual universe contains including duplicates of us. Unlike the actual universe, however, the parallel universe is exactly like Universe A. Finally, in condition (4) a counterfactual indeterministic world, participants were asked to: Imagine scientists discover that the actual universe (the one where you and I live) is exactly like Universe A. Now imagine scientists are suddenly able to observe a parallel universe. The parallel universe contains everything that the actual universe contains including duplicates of us. Unlike the actual universe, however, the parallel universe is exactly like Universe B.

Following this set up, participants were then asked how much they agreed, on a 7-point Likert scale ranging from 1 (completely disagree) to 7 (completely disagree), with one of following randomly assigned statements: (1) If these scientists are right, then the people in Universe [A/B] have free will (2) If these scientists are right, then the people in Universe [A/B] *do not* have free will. Presenting the statement two different ways was to control for question effects on people's response. No significant effect of question type was found so in order to amalgamate these results, levels of agreement for question (2) were reverse coded (i.e., a response of 1 was transformed into a response of 7; a response of 2 was transformed into a response of 6, and so on). After this reverse-coding the results are as if all participants had been asked their levels of agreement whether or not the people in the Universe being evaluated have free will. Thus, higher levels of agreement indicate participants' collective response for there being free will as opposed to not being free will. These results are presented and discussed in §2.3.1.

**2.3.3 Conclusion: Detecting conditionality in the folk concept of free will**

In summary, in accordance with the conditional concept of free will I have presented some new evidence in support of both the weak and strong signal for conditionality. The weak signal for conditionality predicts that people who believe the actual world is indeterministic should straightforwardly respond that there is free will in the actual deterministic world. Consistent with Roskies and Nichols (2008) I found that people who believe that the actual world is indeterministic are unsure whether or not there is free will in this condition. The strong signal for conditionality predicts that people who believe the actual world is deterministic should straightforwardly respond that there is no free will in a counterfactual deterministic world from the perspective of an actual indeterministic world. This time I found that people who believe the actual world is deterministic were unsure how to respond.

While people did not straightforwardly respond in the manner predicted by the conditional concept, I still think this result is evidence in its favour and thus is evidence in favour of the folk being free will compatibilists. That's because I think it would be an amazing coincidence if people who believe the actual world is indeterministic just happen to be unsure how to respond to an actual deterministic world and that people who believe the actual world is deterministic world just happen to be unsure how to respond to a counterfactual world from the perspective of an actual indeterministic world. After all, both these conditions are correctly identified as being important with respect to people's concept of free will, only after it has been identified that our concept of free will might be a conditional concept. Still it is slightly weaker evidence that I would have liked and future longitudinal studies are necessary to confirm that it is a conditional concept that underlies people's responses in this task.

### 2.4. Objection: Two-dimensionalism and counterpossible judgments[26]

Before I conclude this chapter I want to briefly address a few possible objections that you might have about the conditional analysis of our free will concept and the validity of people's responses to certain experimental conditions. First, my conditional analysis of our concept of free will has leaned on two-dimensionalism, which is a purely intensional theory. However, any account of the content of our concept of free will has to be able to model hyperintensional distinctions. In response to this concern, I will very briefly outline a simple extension of two-dimensionalism that will allow us to do just that.

Second, some might worry about the validity of some responses participants were asked to make in my investigation. In particular, you might think that some conditions described metaphysically impossible worlds. Participant responses to metaphysically impossible worlds are unreliable because our intuitions and judgments about free will are not at all concerned with these kinds of worlds. In response to this worry, I argue that even if some things are as a matter of fact metaphysically impossible, we can still imperfectly conceive of them in a way that is sufficient to fix our intuitions and judgments about free will.

#### 2.4.1 Technical issue: Two-Dimensionalism is an intensional theory

Two-dimensionalism, in all its published forms,[27] is a story about logically possible worlds which along one dimension are considered as actual, and along another are considered as counterfactual, but all of which are strictly possible. But any account that hopes to accurately map the content of our concept of free will needs to be able to model hyperintensional distinctions. These are distinctions which are more fine-grained than intensional ones. Thus there are distinctions between worlds that are impossible, and between those that are necessary. Why is this? It's because for all we know the space of accounts of free will is a space in which many accounts are impossible and only the correct one is possible (and indeed necessary). So for all we know (and as some compatibilists suspect) libertarian powers are incoherent and thus logically impossible; perhaps even compatibilism is thought of as impossible by the lights of some libertarians. So for all we know some of the worlds in the simple two-dimensional tables I have presented are impossible. Thus my account is inadequate.

There is a relatively simple fix to this issue: the complete account of someone's free will concept requires multiple two-dimensional table-stacks in the manner of Braddon-Mitchell

[26]Credit and thanks to both David Braddon-Mitchell and Lok Chi-Chan for raising this objection and assisting in developing the reply.

[27]Major accounts of two-dimensionalism include Jackson's generalized account (1998), Chalmers's epistemic account (2004) and Stalnaker's metasemantic account (1978).

(manuscript).[28] The two-dimensional table you should be familiar from this chapter is the contingency table where there are both possible indeterminist worlds with libertarian powers and possible deterministic worlds with compatibilist powers. But on the supposition libertarian powers are metaphysically impossible we will need an additional table. One might think of each table as representing alternative hypotheses about which worlds are possible. There might be a table containing only indeterministic worlds (representing the implausible view that indeterminism is necessary), another table containing deterministic and indeterministic worlds, but not worlds with libertarian powers (representing the view that libertarian powers are incoherent) and yet another table which contains deterministic worlds and indeterministic worlds, some of which contain libertarian powers, and so on. Indexing then is a slightly more complex thing: rather than a particular world being the only index, we need to index to a world and a table. The world tells us how the agent takes the actual world to be, the table tells us how they take the domain of what is possible to be. Actuality *simpliciter*, then, is a world-table pair (or just a world if we don't identify worlds across tables). The world is the actual world and the table tells us what is as a matter of fact logically possible.

Of course no such apparatus is required for logically omniscient agents whose *a priori* knowledge will locate them in the correct table. But philosophers are not logically omniscient, much less the folk whose judgments we are testing in our experimental work, which is why when we are forming judgments about different metaphysical positions we need to adapt intensional theories in a hyperintensional direction. We would not, of course, have to do this if metaphysical contingentism were true (the view that different views in metaphysics are logically possible, rather than competing accounts of what is necessary) but this is a minority view: and of course the difference between that view and the orthodoxy is itself hyperintensional!

So a complete account of our concept of free will requires considering people's judgements not just on the supposition that things are the way they are contingently, but also on the supposition that things are the way they are necessarily as well. This is not a problem for me, and it can be accommodated by simply extending the analysis and methods I developed in this chapter.

### 2.4.2. Response issue: Counterpossible judgments

Let's suppose for the moment that libertarian powers are metaphysically impossible, which I suspect many compatibilist theorists do think. Even if it turns out that libertarian powers are metaphysically impossible, there is no reason to think that people cannot form intuitions or render judgments about them, of the sort that we observe in experimental philosophy, and, for

[28]This view has been called *three-dimensionalism* (Braddon-Mitchell, personal communication; though see Braddon-Mitchell, 2009).

that matter, philosophical debate. For, it seems to me, even if the worlds that we are considering with libertarian powers are metaphysically impossible, they are still conceivable, albeit in an imperfect matter. All that matters, though, is that what we are imperfectly conceiving of is sufficient to play a role in fixing our libertarian intuitions and judgments regarding free will.

But why think that this is right? Well, for one, different conceptions regarding what powers are necessary for free will, including libertarian powers, must be conceivable because philosophers and the folk are divided over the metaphysical status of the powers that are necessary for free will. Further, for each candidate theory about those powers, whether it's libertarian or compatibilist, there are sincere and thoughtful people who believe that's what our free will is like actually. In order to be able to sincerely and thoughtfully think something is actual, entails, in my mind at least, conceivability of a strong enough sense to explain free will intuitions and judgments. What's more there is no reason to think that what is going on in the case of theorists is not also occurring in the folk and so can explain people's intuitions and judgments in those experimental conditions that may have originally concerned us.

You might still wonder how we can mentally conceive of things like libertarian powers if they are metaphysically impossible. I won't spend long on this issue, as it's tangential to the current thesis, but will make one suggestion. When we conceive of things like libertarian powers we form a mental model that is not constrained by our explicitly held philosophical views. For example, my intuitions and judgments regarding libertarian powers in a world might be driven by a mental model that contains people with the apparent power to control indeterministic outcomes. What's important is that the inputs to this process are not a bunch of explicitly held philosophical theories and that is why even philosophers who think libertarian powers are metaphysically impossible can in some imperfect sense still conceive of them. To be more precise, we might imagine a world that contains elements like those libertarians say the actual world contains, even while still taking the official view that such things are metaphysically impossible. And in a world like that, the libertarian powers we would possess make it possible to exert control over indeterministic outcomes.

### 2.4.3. Conclusion: Two-dimensionalism and counterpossible judgments

In summary, here I have briefly addressed a few possible objections that you might have about the conditional analysis of our free will concept and the validity of people's responses to certain experimental conditions. First, the conditional analysis of our concept of free will leans heavily on two-dimensionalism, which is a purely intensional theory. But any complete account of our concept of free will needs to be able model hyperintensional distinctions, thus my account is inadequate. However, this is simply not an issue for me as these hyperintensional distinctions can

be modelled by extending the analysis and methods developed in this chapter to include more tables.

Second, some people might think that people's responses to metaphysically impossible worlds are unreliable because they think our intuitions and judgments about free will are not at all concerned with these kinds of worlds. In response to this worry, I argued that even if some things are as a matter of fact metaphysically impossible is seems as though we can still imperfectly conceive of them in a way that is sufficient to fix our intuitions and judgments about free will. Furthermore, something like this must be the case as there are sincere and thoughtful people on both sides arguing over the metaphysical status and necessity of various powers for free will.

### 2.5. Conclusion: Folk free will in two-dimensions

In summary, in this chapter I have set about the task of arguing that the folk concept of free will is compatibilist, albeit of a conditional kind. A conditional concept is a single concept that has an indexical component which fixes what is necessary for something depending on what the actual world offers up. Applying this thought to the case of free will I have proposed the following conditional analysis of free will:

> If the actual world is indeterministic, and agents have libertarian powers, then these libertarian powers are what free will is and must be. Else, if the actual world is deterministic, and agents have their preferred compatibilist powers, then compatibilist powers are what free will is.

The conditional concept is a compatibilist concept because no matter how things turn out actually to be, to the extent that we possess this concept we will judge that we possess free will. Further, the conditional concept can accommodate the prevalence and tenacity of people's incompatibilist intuitions, which resulted in the orthodox view that the folk concept of free will is incompatibilist. Some observations of free will theorists and undergraduate students suggests that there might be a connection between what people think the actual world is like, and their view about the nature of free will. This means that it is essential to track people's beliefs about the actual truth of determinism in order to know whether they have an incompatibilist concept of free will, or a compatibilist concept of free will of a straightforward or conditional kind.

There is a clear data signal in experimental philosophy of people responding as though they possess a compatibilist concept of free will and as though they possess an incompatibilist concept of free will. However, any differences in results that purport to show the folk are compatibilists and those that purport to show that the folk are incompatibilists are superficial

and due to differences in methodological set up. These differences in methodological set up are correctly viewed as being important when the possibility of the folk concept of free will being a conditional one is identified.

In accordance with the conditional concept of free will I have presented new evidence in support of both the weak and strong signal for conditionality. The weak signal for conditionality predicts that people who believe the actual world is indeterministic should straightforwardly respond that there is free will in the actual deterministic world. Consistent with Roskies and Nichols (2008) I found that people who believe that the actual world is indeterministic are unsure whether or not there is free will in this condition. The strong signal for conditionality predicts that people who believe the actual world is deterministic should straightforwardly respond that there is no free will in a counterfactual deterministic world from the perspective of an actual indeterministic world. I found that people who believe the actual world is deterministic are unsure how to respond. While people do not straightforwardly respond in the manner predicted, I still think that the fact that people are unsure how to respond in the exact conditions predicted by the conditional concept is sufficiently strong evidence to think that the folk concept is in fact a conditional one. If this is right, and I think it is, then the folk are free will compatibilists.

Finally, I rounded off this chapter by briefly addressing a couple of objections you might have about the conditional analysis of our free will concept and the validity of people's responses to certain experimental conditions. Specifically, I outlined how the analysis and methods developed here can be extended to accommodate hyperintensional distinctions in people's concepts, and why we should think that people can imperfectly conceive of metaphysically impossible things in a manner that is sufficient to fix their intuitions and judgments. In the next chapter I address a pertinent issue that faces any account of free action centred on conscious control. That is, evidence from the brain sciences shows us that conscious brain processes might not be causally efficacious with respect to our actions at all. If this is right then by the lights of my account there are no free actions, and so by extension no one has free will.

## Chapter 3

## The challenge of the brain sciences

In this thesis I defend the idea that our actions are free when they are either directly caused by compatibilistically acceptable conscious psychological processes - freedom-conferring mechanisms - or indirectly caused by those same processes via controlled action procedures. So far, I have established that the folk are compatibilists with respect to free will, albeit of a unique kind. If the actual world is deterministic and contains freedom-conferring mechanisms (henceforth FCMs), then the conditional compatibilist folk concept of free will (see Chapter 2) then this is consistent with there being free actions. However, in order for there to be free actions (and hence free will) these FCMs must be causally efficacious. In this chapter I defend this possibility against the following charge: that there is evidence, as a descriptive matter of fact, that our FCMs cause no actions and so none of our actions are free, and no one has free will. Hence, according to which objection, even if our concept of free will is a compatibilist one, nothing actually answers to that concept because our actions are not caused in the right sorts of ways.

There is a growing, but contentious, body of evidence that purports to show that FCMs do not cause actions. Consider the following simple experiment. Participants are asked to watch a clock and report the time shown when they first feel the urge to perform a simple action. Shortly after they have performed the simple action, the timer on the clock stops and the experiment is reset. The key finding is that the onset of neural activity associated with the action occurs well before the time participant's first report having an urge to perform the action (e.g., Libet, Gleason, Wright, & Pearl, 1983; Bode, He, Soon, Trampel, Turner, & Haynes, 2011). Based on these results, many brain scientists conclude that FCMs cannot be the cause of our actions (e.g., Gazzaniga, 2011; Greene & Cohen, 2004; Libet, 1999). After all, if FCMs cause actions, then the time participant's report their urge to perform an action should occur before (or roughly in line with) the onset of neural activity associated with the action.

The target of these brain scientists is the default conscious control account of free action which I introduced in chapter 1. I have called the compatibilistically acceptable version of this default conscious control account: *direct (conscious control) compatibilism*. According to direct compatibilism, an action is free only if it is directly caused by an FCM. As I see it there are two distinct challenges that arise out of the empirical evidence from the brain sciences: the *broad challenge* and *narrow challenge*. According to the broad challenge, evidence from the brain sciences

shows that FCMs cause no actions. In this chapter I will argue that the empirical evidence does not show this at all, and so the broad challenge is wrong.

However, according to the narrow challenge, evidence from the brain sciences shows that FCMs cause only very few of our actions. I think that the narrow challenge is *plausibly* right and is no less problematic than the broad challenge for direct compatibilism. That's because there appears to be a large discrepancy between the few free actions countenanced by the brain sciences, and the number of free actions that we judge there to be in the world (the *discrepancy problem*). If not enough of our actions are free, then we do not have free will. I will argue that we can avoid the discrepancy problem and accommodate many of our free action judgments by acknowledging that most cases that we judge actions to be free, the actions are not free because they are directly caused by FCMs, but free *indirectly* via sub-personal level mechanisms, which those FCMs have influenced in some way: controlled action procedures. Hence the name of my account: *indirect (conscious control) compatibilism*.

Here, then, is the plan for this chapter. To begin, in §3.1, I will present the broad challenge. According to the broad challenge, evidence from the brain sciences shows that FCMs cause no actions and so none of our actions are free. I will argue that the broad challenge is wrong. These arguments occur over the course of §3.2 - §3.5. Specifically, in §3.2 and §3.4, I will address some important empirical evidence from electroencephalography (henceforth EEG) and functional magnetic resonance imaging (henceforth fMRI) research that purports to show that FCMs do not cause actions, and I will argue that they don't show this at all. In §3.3 I will discuss the manner in which participants might be performing these experiments and will suggest we have good reason to think that actions might be being caused by **action procedures**: a special kind of sub-personal level mechanism which I will explain shortly. In §3.5, I will reinterpret the role played by neural activity in the prefrontal cortex in a manner that is consistent with FCMs causing actions in these studies. In §3.6 I will present the narrow challenge and discrepancy problem, and will show that the discrepancy problem vanishes if can we accept that actions indirectly caused by FCMs via controlled action procedures are free. In §3.7 I will conclude.

### 3.1. The broad challenge

Let's begin; here is the broad challenge presented as an argument:

Broad Challenge

1) [Direct compatibilism] An action is free only if it is directly caused by an FCM.
2) [Empirical premise] Evidence from the brain sciences shows that FCMs do not cause actions.

3) [Conclusion] Therefore, there are no free actions (from 1 and 2).

The argument is valid. Let me briefly describe each premise and how they fit together. [Direct compatibilism] is straightforward, as according to direct compatibilism an action is free only if it is caused by an FCM. If compatibilistically acceptable conscious psychological processes do not cause actions, then according to direct compatibilism (and indirect compatibilism for that matter) none of our actions are free.

[Empirical premise] requires substantial unpacking and is the main focus of this chapter. How do researchers in the brain sciences substantiate the claim that FCMs do not cause actions? Before examining any one study in particular, here is the general schema of how current experimental paradigms and their evidence are taken to support [Empirical premise]. In any given experiment, people perform some simple task while a concurrent recording of neural activity is taken. Researchers then compare the timing of the earliest recorded signs of neural activity associated with an action, with people's self-reports of when they were first aware of the action that they will perform (for example, see Libet, 1985; Soon, Brass, Heinze, & Haynes, 2008; Fried, Mukamel, & Kreiman, 2011). The general finding is that the earliest signs of neural activity associated with an action occurs well before people report being aware of that action. According to these researchers, if FCMs caused actions, then the time people report being aware of the action should occur before (or at least roughly in line with) the onset time of neural activity associated with that action. As this is not the case, FCMs cannot be what caused the action, and so the action is not free.

I will return shortly to describe some of this empirical evidence in more detail, and begin applying some pressure to [Empirical premise]. My concern, briefly for now, is that while it's plausible that this line of evidence successfully generalizes to *many* of our actions, there is no reason to suspect that it generalizes to *all* of our actions. As a result, Broad Challenge will need to be amended if it is to successfully challenge free action and free will, but much more on that later on in this chapter.

Finally, consider [Conclusion] for a moment. If FCMs do not cause actions, how was it that we ever came to believe they caused actions in the first place? Many brain researchers (such as those cited earlier) at this point follow Daniel Wegner, who in the *The Illusion of Conscious Will* (2002) famously argued that all cases of our sense of agency are cases of illusory agency.[29] On Wegner's account illusory agency is created and maintained by our own causal detection mechanism which misclassifies the correlation between FCMs and actions as a causal relationship. Instead, the real reason FCMs and action production are correlated is because they

[29]See also Nahmias (2002) who has described in detail, but not endorsed, Wegner's position.

share an unconscious common cause. Although FCMs and action production occur close together in time, they are in fact entirely independent from one another, which means that none of our actions are free. As it so happens, I do not believe that the current state of the evidence is able to support the conclusion that FCMs never cause actions. For now, though, let's examine some of the purported evidence for the [Empirical premise] of the broad challenge.

### 3.2. The [EEG premise] of the broad challenge of the brain sciences

Perhaps the most well known investigations into the relevance of FCMs for actions have been those performed by Benjamin Libet and colleagues (1983; 1985; though see also 2004). For now, I focus my attention on some broad findings around Libet-style paradigms, especially those with a concurrent EEG recording. In §3.4 I will address some more recent experimental work using Libet-style paradigms with fMRI. Although Libet's original paradigm and analyses are now dated, it's still the basis of, and principle motivator for, many contemporary scientific investigations into free will. However, while contemporary results still appear to show that FCMs play no role in many of our actions, I will argue over the next few sections that there is very little reason to suspect that these results generalize to all our actions.

#### 3.2.1. The Libet paradigm and its results

Here is the basic Libet paradigm. During a standard experimental trial participants watch a clock face that has a spot revolving around it. The participant's task is to report where the spot was located on the clock face when they first felt the urge to flex their wrist. Shortly after they have flexed, the clock stops, and the experiment is reset. The task is performed while a concurrent EEG recording is taken, with specific interest in an electrocortical component of neural activity known as the readiness potential (henceforth RP; for example, see Deecke, Grözinger, & Kornhuber, 1976). The RP is a highly robust electrocortical component of general preparatory neural activity which appears to precede voluntary action.[30] The RP begins up to, and around ~2 s prior to action and neural activity continues to accumulate until the intended action is either performed or abandoned. Libet and colleagues' most notable finding was that participants' reports of when they were first aware of the action they would perform, occurred just prior to action, but after the onset of the RP. That is, they observed the onset of preparatory neural activity for action well before the participants' reported being aware of their action. On the basis of these results, the experimenters concluded that FCMs are not causing our actions. After all, to

[30]The exact functional significance of the RP is contentious, and has been described as reflecting neural activity in the supplementary motor area (SMA), pre-SMA, cingulate motor area and anterior cingulate cortex (for example, see Cunnington, Windischberger, Deecke, & Moser, 2002; 2003; Praamstra, Stegeman, Horstink, & Cools, 1996; for review, see Shibasaki & Hallett, 2006). In this thesis, I make no attempt to settle this matter.

these experimenters, at least, if FCMs are causing actions, then participants' should have reported an awareness of their action before (or at least in line with) the onset of neural activity for that action.

While the general pattern of results remains the same, it is worth mentioning that EEG researchers have also taken interest in a special form of the RP, the lateralized readiness potential (henceforth LRP; for example, see Eimer, 1998; Haggard & Eimer, 1999; Vaughan, Costa, & Ritter, 1968). Unlike the RP, the LRP is a purely motoric component that presents as greater neural activity over the motor cortex contralateral to the side of the action. Further the LRP only begins up to, and around ~800 ms prior to action. Of course, there are numerous other, plausibly important, electrocortical components as well. For some examples, you can see the components extracted out of my own investigations in Appendix A.

### 3.2.2. Basic critiques of the Libet-style paradigms and EEG results

It is important to acknowledge that there are plenty of good general critiques of Libet-style methods, and the appropriate functional interpretation of electrocortical components such as the RP.[31] For example, you might be concerned about whether it is even possible for participants to reliably and accurately report timings of their **sense of agency** (for example, see Banks, & Isham, 2009; Lau, Rogers, & Passingham, 2007). By *sense of agency*, I don't mean anything technical, it's simply a useful piece of terminology used by these researchers to refer to someone's self-reported awareness of the action that they will go on to perform. Alternatively, you might wonder whether the RP in the Libet-style paradigm is associated with the action that will go on to be performed at all (for example, see Miller, Shepherdson, & Trevena, 2011; Schurger, Sitt, & Dehaene, 2012). For argumentative purposes, I will set aside some of these more general critiques. That is, I will simply assume people can make reliable and accurate introspective reports, and that the RP, and related electrocortical components, are in fact representative of some form of preparatory neural activity for action production. Although these critiques might be right, it is far from clear that they are, so it is worth seeing how far we can get assuming that they are not.

Another common response to this evidence is to do something similar to what I will be doing, and challenge the generalizability of these findings on the grounds that they involve a very special kind of action (for example, see Tallis, 2011; Mylopoulos & Lau, 2014). It is not hard to motivate this response, especially once you see the kinds of task instructions that Libet and colleagues (and many experimenters to come) presented to participants:

[31]For some more recent general critiques, see Roskies (2010), Smith (2011), and Mele (2014).

> "[…] let the urge to act appear on its own at any time without any preplanning or concentration on when to act" (1983; pg., 625).

For many, this instruction is utterly mysterious and it remains unclear how a participant would go about implementing such an instruction. Furthermore, if this kind of instruction could be followed accurately, then it would rule out the relevance of these experiments for determining whether or not FCMs cause actions. After all, the experimental instruction I just quoted appears to be an instruction to not use FCMs at all in the experiment! At best these results would show that some of our actions are not free, but that's hardly an interesting result as no one ever thought that all our actions were free.

If this is what I thought was occurring in these tasks, then I might conclude that there is already enough reason to think these results fail to generalize. However, while I don't think these results generalize, it's not because I think they reflect some rare or special form of action. In fact, as I will suggest in the next section, I think they probably reflect a perfectly mundane and common kind of action: actions caused by **action procedures**. I will explain what I mean by this shortly; but briefly an action procedure is a sub-personal level mechanism that causes actions. Action procedures will become important soon as I think that the actions caused by a certain kind of action procedure, those influenced by FCMs, (i.e., controlled actions procedures) are free.

### 3.2.3. Conclusion: The [EEG premise] of the broad challenge of the brain sciences

Let's do a quick recap of what I have discussed so far. During a Libet-style paradigm participants report the time at which they first experience a sense of agency for some basic action. The results of these experiments show that the neural activity for the action occurs well before participants' earliest reported sense of agency. On the basis of these results some researchers have concluded that FCMs cannot be what are causing our actions. They further conclude that there are no free actions, because they assume that the results of these experiments generalize to all our actions. If FCMs *never* cause actions, then of course there are no free actions (or free will for that matter). There are many general critiques of the methods and results of these kinds of studies; for example, what exactly does the RP realise in the context of action preparation and performance? Further, many people think that these results don't generalize because the kinds of actions being examined are rare or unique. While I agree that the results of these experiments fail to generalize to all our actions, unlike others I think the actions performed in Libet-style paradigms might reflect the vast majority of our actions: namely, actions caused by sub-personal level mechanisms, or what I have called action procedures.

### 3.3. Participant performance in a Libet-style paradigm

So, what are participants doing during a Libet-style paradigm? Consider Alfred Mele's (2014) description of his experience and approach when performing a Libet-style experiment:

> "At first, I waited for conscious urges to flex to pop up in me so I'd have something to report when it was time to make the *consciousness* report. I waited until I was pretty sure the urges weren't just going to pop up on their own. I wondered what to do, and I hit on a strategy. I decided that I would just say "now" to myself silently, flex my wrist in response to the silently uttered cue, and then, a little later, try to report where the hand was on the Libet clock when I said "now". (pg. 14)

While Mele's experience is of course his own, and different people will report different experiences, there are two important points to take away from his report. First, actions simply do not appear out of nowhere. For example, it is hard to imagine how participants could bring about wrist flexes in the experiment without any thought for the requirements of the experiment they are participating in. Second, actions that apparently appear out of nowhere are in fact caused by some action procedure of the agent. I will expand on this shortly.

But first, what is an action procedure exactly? At the beginning of the thesis I introduced the idea that an action procedure is a sub-personal level mechanism that causes actions, or influences other action procedures, both in a contained and reliable fashion. So imagine for the moment a simple fist-bumping action procedure that produces a fist-bumping action. The action procedure is contained in the sense that it will only produce actions of the fist-bumping kind (contained) and will rarely fail to produce hand-waving actions whenever it is activated (reliable). If an action procedure is completely independent of FCMs, then the actions it causes are unfree. However, I think that many actions procedures are influenced by FCMs, and, what's more, I think that action procedures that have been influenced by an FCM cause free actions, albeit indirectly.

I have called an action procedure that has been influenced by a FCM: a **controlled action procedure**. The simplest kind of influence an FCM can have on a action procedure is **creation**. For example, an agent might create a waving action procedure that causes them to wave whenever someone waves at them. Alternatively, an agent might create an action procedure to automate their actions while performing a Libet-style paradigm. This controlled action procedure (henceforth CAP) would reliably cause them to flex their wrist in a timely manner some time after the start of an experimental trial has been cued. Exactly, what sorts of connections between an FCM and an action procedure count as *influence* will be discussed in

detail in Chapter 4 (§4.1). For now, I will stick with creation as a paradigm case in which an FCM influences an action procedure.

It's worthwhile to note that I don't think that action procedures and CAPs are natural kinds which are identifiable by the low-level brain sciences. That is, I do not think that action procedures and CAPs are some special kind or patterning of neural activity. Instead, action procedures are sub-personal level mechanisms that are defined functionally as whatever those things are that cause actions in a contained and reliable fashion, and when they have been influenced by a FCM - CAPs - cause actions which are indirectly free according to indirect compatibilism.

If this general thought is correct then the reason Libet-style results fail to generalize is not because they target some special or rare kind of action. The opposite is true. It's because they in fact fail to generalize to a special and relatively rare kind of action: the act of an FCM creating (influencing) some action procedure. For, I will suggest, it is the fact that FCMs are involved, in an appropriate manner, in the creation (influencing) of action procedures, which is responsible for many of our free actions, and hence for vindicating the thought that we have free will. So while I also think there are some cases where FCMs *directly* cause actions, the focus of my discussion will be on the act of creating (influencing) action procedures.

### 3.3.1. Unpacking participant performance

Let's start elaborating. First, as Mele describes, simply waiting for urges to flex one's wrist seems a non-starter in order to successfully complete a Libet-style experiment. For one, there is no guarantee that the required urge would ever arise over the course of the experiment - if at all - if there were no CAP in place to cause the action. This issue is compounded by the fact that these experiments rely on there being numerous successfully completed trials. At the very least, successful experimental completion requires some CAP to cause some wrist flex at some time during each experimental trial. As far as I am aware, there has never been a reported case of some poor participant spending an experiment simply sitting and waiting, with no success, for the urge to flex their wrist to arise on its own!

Of course, I do not think that creation (influencing) requires that people make the conscious decision to create a specific CAP, or bring it about that one CAP, as opposed to another, activates and causes actions. Instead, I think that the participant's conscious decision to participate in the experiment is realized by the creation (influencing) of an action procedure, which automates and causes the participant's actions for the duration of the experiment. There will be some relatively rare cases where someone does create (influence) a very specific action procedure via conscious control. For example, expert musicians might create (influence) an

action procedure to cause a very precise and specific set of actions via conscious control. I recall performing finger dexterity exercises on the piano which began by consciously controlling individual finger movements in a precise manner over and over again until such time as these precise finger movements became entirely automated. Expert sportspeople very likely do something similar in their training.

Second, while Mele describes the CAP he adopted to perform the experiment, there are many kinds of CAPs participants might adopt. This is especially true of Libet-style paradigms for a number of reasons. For example, performance is not assessed in such a manner that it makes sense to judge the actions caused by one CAP as being better, or worse, than those caused by another CAP. There are never any incentives for performance, so there is no motivation to adopt one kind of CAP as opposed to another. Finally, there was no feedback provided for performance. Aside from the internal motivation a participant might have, perhaps curious to try out different CAPs, once an action procedure had been created (influenced) there was no reason for a participant to adjust it on account of the actions that were being caused. Of course, not all contexts will be like this. Different CAPs might result in differential performance that can accurately be assessed as better, or worse, on account of the actions they cause. Further, our CAPs will often be adjusted and improved according to the feedback we receive for the actions we perform.

**3.3.2. Controlled action procedures and our sense of agency**

So our main question is whether the actions caused by our CAPs are free. While, 'yes' will be my answer shortly in this thesis, the direct compatibilist has to answer '*no*'. As already described, Libet and many other researchers have presented evidence that our sense of agency for an action caused by a CAP only occurs after the onset of preparatory neural activity for that action. However, it also appears that our sense of agency for these kinds of actions is not even associated with the preparatory neural activity (for example, see Haggard & Eimer, 1999; Schlegel, Alexander, Sinnott-Armstrong, Roskies, Tse, & Wheatley, 2013). Neuropsychology has provided valuable information about the neural source of our sense of agency for these kinds of actions. For example, Sirigu and colleagues found that patients who suffered parietal lesions due to stroke, while able to report the onset time of their physical movements, were not able to report the onset time for sense of agency (Sirigu et al., 2004)[32]. Meanwhile, Desmurget and colleagues (Desmurget, Reilly, Richard, Szathmari, Mottolese, & Sirigu, 2009) applied direct electrical stimulation to seven patients undergoing awake brain surgery. When direct simulation was applied to premotor regions, it triggered overt movements in patients. Despite this, patients

[32]However, this result was not replicated by Larfargue and Duffau (2008). They examined patients undergoing surgical resection of their inferior parietal lobule due to slowly developing brain tumours.

firmly denied they had made a movement. In contrast, when direct stimulation was applied to parietal regions, patients reported believing they had moved, even though no movement was recorded. Our sense of agency for these kinds of actions appears to be associated with some completely different processes, apart from action production.

Still, it's important to remember these results only appear to reflect one kind of action we might have thought was free - actions caused by our CAPs. And while, I think this kind of action does not count as being free according to direct compatibilism, this result alone is not enough to show there are no free actions and no free will. In order to show that, we would also require evidence showing that creating (influencing) action procedures is associated with the same patterning of neuroimaging results as the actions they cause. After all, if these neuroimaging results generalize to the creation (influencing) of action procedures, then, by my lights, they will have generalized to all our actions!

In the next section we are going to consider some research carried out using fMRI. These researchers have claimed the distance in time between the onset of preparatory neural activity and our sense of agency might be as much as 10 s (for example, see Bode, He, Soon, Trampel, Turner, & Haynes, 2011; Soon et al. 2008; Soon, He, Bode, & Haynes, 2013; though see also Schultze-Kraft et al., 2016). If they are correct, it could be suggestive evidence that our sense of agency for the creation (influencing) of our action procedures, which plausibly operates over a much bigger time-scale, faces the same issues as our sense of agency for the actions they cause. If so, this would undermine my contention that we can freely create (influence) action procedures.

### 3.3.3. Conclusion: Participant performance in a Libet-style paradigm

In summary, it's very unlikely that participants performing Libet-style paradigms are performing the experiment in the manner researchers think they are. There is simply no guarantee that the urge required by the task would ever arise on its own during an experimental trial, let alone repeatedly over the course of the experiment. Instead, it's plausible that participants are creating (influencing) an action procedure. The CAP in this case would be a remarkably simple sub-personal level mechanism that would unconsciously cause a simple wrist flex whenever the onset of an experimental trial was cued. Actions caused in this manner are still associated with our sense of agency, but this sense of agency is not associated with causing the action itself. Further, while actions caused by CAPs are not free, according to direct compatibilism, they are free, albeit indirectly, according to my own account of free action: indirect compatibilism. Indirect compatibilism will be the major focus of Chapter 4.

### 3.4. The [fMRI premise] of the broad challenge of the brain sciences

So far in this chapter I have been discussing the [Empirical premise] of the broad challenge. The premise states that the current results of investigations in the brain sciences have shown that FCMs do not cause actions. Up to this point I have primarily focused on results obtained from research using EEG; now I will shift my focus to results obtained from research using fMRI. Looking forward, my judgement will remain unchanged; these results fail to generalize to all our actions. This means that they cannot rule out there being some free actions in the world. Further, it is possible to interpret some fMRI results as consistent with FCMs causing actions. I will not argue that this is the correct way to interpret these results, only that it's one option.

Researchers employing fMRI have employed experimental methods similar to those developed by Benjamin Libet. For example, Soon, et al. (2008; though see also Bode, et al. 2011) instructed participants to fixate on a stream of letters, and press a button with either their left-or-right hand when they first felt the urge to do so. Once they had pressed a button, participants then selected the letter that was present on the screen when they first felt the urge to perform that action. These researchers were interested in how much predictive information for a participant's action was present in various brain regions before a participant's reported sense of agency for that action.[33] What they observed was quite remarkable; patterns of neural activity in brain regions such as the frontopolar cortex, medial prefrontal cortex, and substantial areas of the parietal cortex, were predictive above chance (~60%) of the action a participant would perform. Further, this was up to 10 s before participants first reported their sense of agency. Similar to Libet and others, it was concluded that FCMs could not be causing our actions, as the onset of preparatory neural activity for the action occurs well before our reported sense of agency for that action.

#### 3.4.1. Contemporary advancements in the Libet-style paradigm

Before I describe some other plausible explanations of what might be occurring in these studies, I want to draw attention to two important advancements brought about by this kind of research. First, it is natural to see these results as a natural extension of some of the original findings using the RP and LRP. As you advance closer to the time participants report a sense of agency for the action they will perform, predictive neural activity can be observed in various other brain regions as well. This includes the supplementary motor area and primary motor cortex, important source

[33]While studies conducted using fMRI will be the focus here, the analysis technique being employed, multivariate pattern classification has been applied to other kinds of neuroimaging data as well. For example, see Bode, Sewell, Lilburn, Forte, Smith, & Stahl, 2012, for the analysis as applied to EEG data. I will be assuming for the purposes of discussion that preparatory activity observed using fMRI actually represents neural activity in those regions. There is some evidence this is not the case (for example, see Sirotin & Das, 2009).

locations for electrocortical components such as the RP and LRP (for review, see Shibasaki & Hallett, 2006). However, these regions only became predictive, at best, just prior to participants' reports. This suggests that the amount of time between the onset of neural activity and the sense of agency for some actions is greater than we originally thought. Second, these researchers have begun moving on from investigating simple motor movements, and have begun investigating (relatively) more complex actions as well. For example, Soon, et al. (2013), using a method like the one described previously, required participants to perform simple addition or subtraction, rather than simply pressing a button. While there was the additional observation of neural activity in a brain region associated with arithmetic, there were no differences in the predictive brain regions observed in this study.[34] This suggests these regions are not just important for voluntary motor actions, but voluntary actions more generally.

**3.4.2. My preliminary assessment**

Ultimately, my own thoughts about what is occurring in these kinds of studies remain unchanged from what I have said previously. This should come as no surprise, as the most substantive changes between these studies and those described earlier regard the neuroimaging and analysis techniques employed. The experimental task itself is essentially unchanged. As a consequence, there should be no substantive changes to participant performance. Even if participants are unconsciously causing actions in these tasks, it does not follow that all actions are caused in this manner. No matter how accurately these results might reflect the actions caused by our CAPs, there is simply no reason to suspect they generalize to reflect the act of creating and influencing the action procedures themselves.

If participants are using a CAP, then the earliest signs of neural activity in these studies might realise the creation (influencing) of that action procedure. *Prima facie* this suggestion is also consistent with the patterns of neural activity observed; as neural activity in the frontopolar cortex has been associated with plan implementation (for example, see Bunge, Kahn, Wallis, Miller, & Wagner, 2003; Sakai & Passingham, 2003; 2006), meanwhile the medial prefrontal cortex, in conjunction with the parietal cortex, is an important part of a fronto-parietal network associated with stable and novel task control (for example, see Braver & Barch, 2006; Cole, Reynolds, Power, Repovs, Anticevic, & Braver, 2013; Dosenbach et al. 2007). However, as with most frontal regions, there is no clear consensus on the common functional roles they fulfil for task performance. In §3.5 I will go on to describe some further roles neural activity in these regions have been associated with.

[34]In this case, the angular gyrus (for example, see Dehaene, Piazza, Pinel, & Cohen, 2003; Grabner et al. 2009; Stabescu-Cosson et al., 2000; for contra evidence, see Van Harskamp, Rudge, & Cipolotti, 2002).

Still, if we assume participants' performance in a Libet-style experimental protocol is the result of a simple CAP, I believe we can make good sense of these researchers' original claim. That is, it seems plausible to think participants who are using a CAP might be unconscious of the process by which actions are caused by that CAP. The obvious cost of this explanation is it means these results cannot generalize to all our actions, as it remains open that FCMs might still play a role in creating (influencing) the action procedure itself. It is then a further empirical question whether the creation (influencing) of action procedures is associated with the same patterning of neuroimaging results we observe for actions caused by CAPs. While I think this is the right explanation, it's not what these researchers had in mind. For now, I will set my own thoughts aside and address two ways they might have interpreted these results.

#### 3.4.3. Conclusion: The [fMRI premise] of the broad challenge of the brain sciences

To summarise to this point, many contemporary researchers using Libet-style paradigms have begun to utilize fMRI. These researchers have also made two major advances. First, EEG researchers were principally interested in electrocortical components around the time the action was about to be performed (i.e., RP and LRP). fMRI researchers, on the other hand, have begun identifying predictive neural activity in diverse regions well before the action occurs. The amount of time between the onset of neural activity for an action and our sense of agency for that action appears to be far greater than was suggested by the earlier EEG studies. Second, the shift away from overt motor actions has allowed researchers to generalize and find similar patterns of predictive neural activity for a diverse range of actions such as performing basic arithmetic. However, the core of these tasks remains unchanged, which means that participants might still be performing these tasks by creating (influencing) a simple action procedure. If participants are using a CAP, then participants might very well be unconscious of the process by which actions are caused in these experiments. What's missing is evidence that participants are also completely unconscious of the creation (influencing) of their action procedures.

### 3.5. Reinterpreting the role of the prefrontal cortex

Here are two ways these results could be interpreted so as to be consistent with the conclusion that participants are unconscious of the process by which actions are caused in these experiments. First, patterns of neural activity observed in these brain regions might represent the onset and ongoing preparatory neural activity for an action, and this neural activity is not associated with any FCM. Second, patterns of observed neural activity represent the action to be caused. While participants think their FCMs cause their action, they are mistaken. The action they will perform is completely determined by the predictive neural activity or that predictive

neural activity in conjunction with some further unconscious neural activity. In what follows, I will discuss each possibility and suggest that, even if they are consistent with the data, there is nothing in the current results which commits us to believing either case is actual. The upshot of this is that there's no reason to accept the [Empirical premise] of the broad challenge. FCMs *plausibly* do cause some actions and so some of our actions are free.

### 3.5.1. Interpretation one: No FCMs?

Let's assume for the moment these researchers are correct: that the onset and ongoing preparatory neural activity observed in brain regions is not actually associated with FCMs. What are participants thinking and doing in the 10 s between the onset of the predictive neural activity and the earliest reported sense of agency for that action? Before, I suggested one way to make sense of this; the observed neural activity is associated with an action procedure being created (influenced) to cause actions in the experiment. This leaves open what else these participants might be doing. Perhaps they are ruminating on what to do in the evening, the episode they are up to on their favourite Netflix series, and so on. However, these researchers believe their results generalize to all our actions, so this cannot be how they interpret their results. In order to figure out what participants could be thinking and doing in any given experimental trial, let's see what other activities prefrontal regions are associated with. For readers that are interested, §3.5.1.1 contains some empirical evidence to support my proposal. I will return to discuss parietal regions in Chapter 4 (§4.2). Reflecting on just prefrontal regions is enough for my purposes here.

Earlier I stated that there is no consensus on what common functional role prefrontal brain regions play in task performance. This is important as it means there is no reason to think the neural activity observed here *must* realise unconscious preparatory neural activity. Instead, it's just as plausible to think that the neural activity observed in the frontopolar and medial prefrontal cortex is a candidate realiser of some FCMs. I have already described one task which has been associated with neural activity found in the prefrontal cortex: plan implementation. Further examples will be provided shortly, §3.5.1.1, for those who are interested. While this neural activity is not associated with the sense of agency for the action performed, it could still be associated with our sense of agency for the creation (influencing) of action procedures. Let's call this subpart of our sense of agency, our **sense of deliberative agency**. Provided our sense of deliberative agency is correct, which is to say it's associated with our FCMs creating (influencing) action procedures, then we appear to have all that is required to vindicate the existence of some free actions. At the very least, there is nothing which means the current empirical evidence *must* be interpreted as threatening this view.

*3.5.1.1. No FCMs: Evidence.* In what follows I will very briefly describe some empirical evidence to support my proposal that neural activity in the frontopolar cortex and medial prefrontal cortex is a candidate realiser of some FCM. This is important as it means that the neural activity in these regions in Libet-style paradigms might sometimes realise an FCM causing a free action. In the context of these kinds of experiments, I think the most plausible candidate is the creation (influencing) of an action procedure to automate participants' performance.

In addition to plan implementation, neural activity in the frontopolar cortex has been associated with our capacity for autobiographical and prospective memory (for example, see Reynolds, McDermott, & Braver, 2006; Okuda et al. 2003; Burgess, Gonen-Yaacovi, & Volle, 2011), as well as counterfactual, relational and abstract reasoning (for example, see Boorman, Behrens, Woolrich, & Rushworth, 2009; Bunge, Helskog, & Wendelken, 2009). Meanwhile, neural activity in the medial prefrontal cortex has been repeatedly associated with decision making, particularly involving risk and reward (for example, see Bechara & Damasio, 2005; Fellows, 2007). This list is far from exhaustive.[35]

Once again, what's important for my purposes is the fact that neural activity in these prefrontal brain regions is closely associated with our capacity for future planning, complex reasoning, and sensitivity for risk and reward: cognitive processes which closely reflect the FCMs described by compatibilists.

**3.5.2. Interpretation two: Causally inefficacious FCMs?**

Of course, researchers interested in free will acknowledge the proposal I outlined, and the variety of task roles these brain regions are associated with. Yet, they still conclude that the onset and progression of predictive neural activity through various brain regions are associated with unconscious processes. For example, Soon, et al. (2013) when discussing the roles of the medial frontopolar cortex, posterior cingulate/precuneus say:

> "Both regions that encoded the content of the decision ahead of time have also been implicated in tasks involving both prospective memory, and imagining both the past and future. A possible interpretation of the current results is that these regions were involved in unconscious preparation for actions in the near future." (pg. 6220)

I have already suggested that these results are consistent with the conscious preparation of our actions. However, this quote makes clear another way our actions could be unconsciously

[35]Other important functional roles for the frontopolar cortex includes integrating together the results of disparate cognitive mechanisms (for example, see Ramnani, & Owen, 2004) and the resolution of cognitive uncertainty and conflict (for example, see Badre & Wagner, 2004; Koechlin, Corrado, Pietrini, & Grafman, 2000; Yoshida & Ishii, 2006). Other important functional roles for the medial prefrontal cortex includes, error detection and conflict monitoring (for example, see Botvinick, Cohen, & Carter, 2004; Holroyd, Coles, & Nieuwenhuis, 2002), as well as incentivise-based learning (for example, see Rushworth, Noonan, Boorman, Walton, & Behrens, 2011). There are still more examples that might be listed.

caused. While there are some FCMs, those FCMs do not contribute to the action. Nor do any other FCMs contribute to the action. Instead, the action is completely determined by the predictive neural activity. Or, if the action is not completely determined by the predictive neural activity, what determines the action is not an FCM but further unconscious neural activity. As I said earlier, there is no reason to think additional preparatory neural activity must only represent unconscious preparatory neural activity. But, insofar as our future action can be predicted from patterns of neural activity in prefrontal brain regions, it's highly unlikely that this prediction is being performed by decoding a completely determined action. I will come to this point shortly. Moreover, the success of our predictive capabilities does not require there be a completely determined action encoded at the onset of predictive neural activity.

First, predictive success in these studies does not require a completely determinable action be encoded from the onset. Nor does it require the action that will be performed be encoded at all. When participants are asked to generate a 'random' sequence of binary responses, as they were in these studies (i.e., perform an action-or-not, left-or-right hand button press, addition-or-subtraction) they switch between responses with probability of 0.6 (for example, see Lopes, 1982; Treisman & Faulkner, 1987; Falk & Konold, 1997). The fact that people's behaviour is not random, but instead can be modelled using a switching-rule, means it's not only possible to predict, above chance, people's actions using the neuroimaging data in these studies, but using the behavioural data as well (Lages, Boyle, & Jaworska, 2013). This means it's plausible that what is being decoded from this earliest predictive neural activity is some abstracted task-rule, or CAP. This would not only explain the predictive nature of this neural activity, but also its substantive failure rate. After all, even if a CAP is causing actions, there are a number of different ways different CAPs might do this.

Second, the most prominent neurobiological model for voluntary action ascribes different functional significance for action depending on its location along an anterior-posterior gradient, that runs from the prefrontal cortex toward the supplementary motor area (for example, see Jahanshahi, 1998; Brass & Haggard; 2008; Zapparoli, Seghezzi, & Paulesu, 2017). Neural activity in anterior regions represents actions in only their most abstracted form, such as in the form of some overall goal or intention. As neural activity progresses posteriorly over time toward motor regions, it is completed in detail. For example, Brass and Haggard (2008) describe the function of the anterior-posterior gradient in terms of the what, when, and whether, of voluntary action. On this account, anterior neural activity is associated with whether to perform an action or not, and posterior neural activity is associated with when to perform the action. In-between the two ends of the gradient what action to perform is represented in varying levels of

detail. For my purposes, all that is required to note is that our actions are not completely determined by predictive neural activity, but instead, appear to be caused iteratively. This leaves plenty of room for FCMs to contribute to and determine what action is caused. (Or whether, for that matter, an action is caused at all).

It should be obvious at this stage that the [Empirical premise] of the broad challenge is mistaken. There is nothing in the current empirical literature that must be interpreted as undermining the existence of some free actions of the kind posited by direct compatibilism. On the other hand, the proposals outlined according to which even these simple tasks might be directly produced by FCMs seem implausible. There is no reason to suspect that FCMs would ever be engaged in the kinds of experimental tasks these participants performed. That's because they are deliberately impoverished of the kinds of things people think matter for free action. For example, there is no need to deliberate, there are no incentives for performance, and there is no reason to shift performance once it has begun. However, just because these kinds of tasks do not rely on FCMs does not mean FCMs don't play a role when in fact they might be required. I have noted throughout this section that I think performance in these experiments reflects the creation (influencing) of an action procedure. And it's still an open question whether the creation (influencing) of action procedures is subject to the same interpretation as actions caused by CAPs. No matter which path you choose, we are left in the position of concluding that the [Empirical premise] of the broad challenge is false. It's very likely FCMs do create (influence) action procedures and so it's very likely FCMs do on occasion cause actions.

### 3.5.3. Conclusion: Reinterpreting the role of the prefrontal cortex

In this section, I discussed two interpretations of researchers' claims that participants are unconscious of their task-relevant preparatory neural activity. First is the claim that predictive neural activity observed in the prefrontal cortex in these tasks only realises unconscious preparatory neural activity. Second is the claim that predictive neural activity in the prefrontal cortex, even if it realises our sense of deliberative agency, might be causally inefficacious with respect to the action that we perform. In response, I have provided evidence that neural activity in the frontopolar and medial prefrontal cortex is a candidate realiser of some FCMs including: future planning, complex reasoning, and sensitivity to risk and reward. Moreover, I have argued that actions are not completely determined by predictive neural activity in prefrontal brain regions but are completed as they project back to motor regions for execution. This leaves plenty of room for FCMs to shape and cause our actions. It should be clear at this stage that the [Empirical premise] of the broad challenge from the brain sciences for free will is not true. With that said, I think these researchers are correct that participants are unconscious of the vast

majority of task-relevant neural activity in these experiments. The best way to make sense of this claim is by appealing to CAPs. Once a participant has created (influence) an action procedure to automate performance there is no need to think about the experiment anymore! In doing so they free themselves up to bring to mind, ruminate on, and plan something else.

### 3.6. The narrow challenge

So far, I have shown there is no reason to think, based on current empirical results, that FCMs *never* cause actions. However, as I have stressed, although these results might fail to generalize, they still might describe most instances of action production (i.e., the actions caused by our CAPs). They only fail to represent select instances of action production where FCMs could matter, (i.e., a few of our actions and the creation (influencing) of the action procedure to cause our actions). With this in mind, here is the narrow challenge:

Narrow Challenge

1) [Direct compatibilism] An action is free only if it is directly caused by an FCM.
2) [Empirical premise] Evidence from the brain sciences show that FCMs cause very few actions.
3) [Conclusion] Therefore, there are very few free actions (from 1 and 2).

The argument is valid. The difference between the broad and narrow challenges of the brain sciences is the strength of Premise (2). While Premise (2) of the broad challenge claimed that current empirical results show that FCMs cause no actions, in the narrow challenge it claims they cause very few actions. So it is compatible with the success of the Narrow Challenge that at least a few of our actions are free. Nevertheless, as I will now argue, this is cold comfort to the compatibilist.

#### 3.6.1. The discrepancy puzzle

*Prima facie* if the narrow challenge is the strongest challenge the brain sciences offers to free action, then the result is good for the direct compatibilist. Further, it might be difficult to see why we should continue to see the narrow challenge as problematic at all. The reason we should is because there is something problematic about the thought that the number of free actions countenanced by the brain sciences is far fewer than the number free actions that we judge there to be in the world. This is what I have called the Discrepancy Puzzle:

Discrepancy Puzzle

1) [From conclusion of narrow challenge] According to the brain sciences there are very few free actions.

2) [*A priori* premise] We only have free will if most of our actions which seem to us to be free are free.
3) [*Plausible* empirical premise] There are very many actions which seem to us to be free.
4) [Lemma] There are very many actions which seem to be free, but are not (from 1 and 3).
5) [Conclusion] Therefore, we do not have free will (from 2 and 4).

Once again, the argument is valid. The first premise of Discrepancy Puzzle is just the conclusion of Narrow Challenge. Since (4) follows from (1) and (3), and (5) follows from (4) and (2), I only need to defend (2) and (3). I'll start by defending (3). Of course, this is an empirical premise. In the end, a full defence of the premise would require a good deal of experimental work. Still, I think the premise is plausible. Reflection on my own intuitions, and those of others, suggests to me that people do not just have a sense of agency when their FCMs are causing actions. Instead, our sense of agency is the default, and in the absence of any defeaters to our free will, we believe we are acting freely. (Where standard defeaters to free action include instances of physical restraint, psychological coercion and manipulation, and being the victim of a nefarious neurosurgeon who feels the need to tamper with your brain). Here's the idea. I might deliberate in the morning over where to purchase my coffee in the morning prior to meeting with my PhD supervisor. However, I don't just feel free during my time spent deliberating. Instead, I felt free as I opened the door, stepped outside, closed and locked the door behind me, walked up the road, petted a cat on the side of the road that I have affectionately named McSnoots, and so on. Without evidence to the contrary, I just have the sense I am free. I think most of us are like this, and hence, I think that we suppose a wide variety of our actions to be free.

Moreover, this supposition would explain why Libet and other brain scientists have thought that the actions they have investigated would be considered free according to the folk concept of free action. If acting freely is as prevalent as it seems, then starting scientific investigations with the *simplest* actions that would be judged free makes good sense for the purposes of experimental control. To be clear, just *how* prevalent our feeling of being free is, and whether or not it's more prevalent than the number of free actions countenanced by the brain sciences is a matter for future empirical research. For now, I hope to have said enough to motivate it as a plausible hypothesis, and apt target of future research.

Now let's consider (2). Recall that on my view, free will refers to the general capacity of an agent to reliably act freely. I think it plausible that our sense of agency is a reliable guide to when we act freely. If most of the time we spend feeling as though we have free will is mistaken, then we are not reliably acting freely, and hence do not have free will. This is true even if there

are select few instances when our FCMs do cause actions. These actions will be free, but because we don't reliably act freely there is no free will itself.

In conclusion then, I think both (2) and (3) are plausible, and hence that if the Narrow Challenge succeeds then we lack free will, even if we perform some free actions. If that is right, then the Narrow Challenge is a significant problem for the direct compatibilist. In what follows I will say something about how I think we ought to respond to the discrepancy problem.

**3.6.2. Determining what free will is**

In order to respond to the discrepancy problem, we need an explanation of why the number of free actions that are vindicated by the brain sciences, appears to be smaller than the number of actions we judge to be free in the world. Once we have explained this, we will see that in fact, in the extended indirect sense that I will defend, that the number of actions that we judge to be free is in fact far closer to the number of actions that are in fact free than first appeared. That is because the free actions vindicated by the brain sciences may pass their vindication indirectly onto many other actions.

There are two ways we can think about how the folk concept of free action determines what free action is. The first way, which I will call **theory-centred approach**, says free action just is whatever is described by our folk theory. I exemplified this approach in Chapter 2 of the thesis. One important reason to perform experimental philosophy investigations is to collect folk judgments in order to reverse engineer the theory they are responding from. It was on the basis of this rationale that I was able to establish that the folk theory of free will is a compatibilist one. While the theory-centred approach tells us that the folk reliably respond according to their theory, there is an open empirical question whether or not there are any free actions in the world. After all, even if the folk judge certain FCMs are necessary for free action, it's a matter of empirical discovery whether or not we are in possession of such mechanisms and whether or not such mechanisms cause actions.[36]

The second way, which I will call the **case-based approach**, says that free action just is, roughly, whatever it is that's in common between all the cases where the folk judge free action is instantiated. Provided there is something that is in common between the cases, the case-based approach virtually guarantees there is free action in the world. Empirically, what's up for grabs is the exact nature of free action. In this part of the thesis I will be adopting the case-based approach in order to answer the discrepancy puzzle, and develop my new account of free action,

[36]The theory-centred approach is helpful even if what you are doing is conceptual engineering as it predicts whether the folk with accept the results of said engineering.

indirect compatibilism. Further, as my empirical focus is on the brain sciences, the nature of free will is described in terms of cognitive and neural systems.

Of course it's possible that the theory-based and case-based approaches not only provide different, but incompatible, analyses of free action. For example, the theory-centred approach might tell us that a free action is an action caused by our FCMs, and we might empirically discover we do not possess these mechanisms. However, the case-based approach could still identify some commonality between all the cases the folk judge to be free. If this turns out to be the case, then two moves are available. If we think the free action analysis provided by the theory-centred approach is the correct one, then we might be moved to become free action (and by extension free will) eliminativists. After all, our theory told us what the viable candidates were for free action, and we discovered there were no such candidates in the world. Conversely, if we think the free action analysis provided by the case-based approach is correct, then we might be moved to radically revise our concept of free action in line with whatever we discover is in common between all the actions in the world we judge to free. I find neither move particularly appealing.

*3.6.2.1. The combined theory-centred and case-based approach to determining what free will is.* Instead, I advocate an alternative approach – we combine the theory-centred approach and case-based approach. According to the combined approach, there is free action just in case the following two conditions are both satisfied:

a) [Case-based approach] There is some phenomenon in common between all the cases the folk judge to be free.
b) [Theory-centred approach] The phenomenon uncovered in (a) satisfies the conditions discovered by experimental philosophy, or else, is something the folk would make principled revisions towards in order to make sure certain cases continue to remain correctly judged as free.

On the combined approach analyses offered up by either the case-based approach, or the theory-centred approach describe the necessary, but not sufficient conditions, for free action. They are only sufficient for free action when jointly satisfied. Each approach has a unique role to play in uncovering what satisfies our concept of free action. The role of the case-based approach is to identify whether or not there are any candidate phenomena to be free actions in the world. And the role of the theory-based approach is to examine whether or not the candidates identified by the case-based approach are good enough deservers to count as being free actions. What I have identified as being in common between all the cases we judge to be free are FCMs. Unfortunately, I have not yet tested whether FCMs are freedom-conferring when they cause an

action indirectly. Indirect compatibilism, as its name would suggest, requires this to be true of our free action theory. *Prima facie*, as indirect compatibilism is a close relative to direct compatibilism, we have some reason to think that indirect compatibilism should also deserve to satisfy our concept of free action. But I do not have the exact answer at this time. I will have much more to say about meta-level issues, such as the status of indirect compatibilism as a theory, the relationship between our theories and judgments about cases of free will and so on, in Chapter 5. Once I have finished presenting my account indirect compatibilism.

### 3.6.3. Conclusion: The Narrow Challenge of the Brain Sciences for Free Will

In summary, even if the broad challenge of the brain sciences for free will fails, there is a nearby successor in the form of the narrow challenge of the brain sciences. Unlike the broad challenge, the narrow challenge is consistent with FCMs sometimes causing actions. While this might appear good for the direct compatibilist, it's a hollow victory. Even if FCMs do sometimes cause our actions, the number of the instances in which they do so is very likely far fewer than the number of instances where we judge ourselves to have acted freely. This difference in the number of cases in the world that are free, and the number of cases in the world we judge to be free, is what I call the discrepancy problem. If the vast majority of cases we judge are ones in which the action was free, are cases in which the action was not free at all, then, I argued, we do not have free will.

In order to determine what free action is, I advocate an investigative approach that combines both case-based and theory-centred approaches. First, we identify what's in common between all the instances in the world that we judge to be free. Second, we submit the phenomena in common between the cases to the folk in order to see whether or not they are important for our free action, or whether the folk could come to accept that the phenomena are sufficient to ground there being free action, in order to secure certain cases in the world as being free. My own account of free action, indirect compatibilism, requires that the folk judge instances where our FCMs *indirectly* cause our actions count as free. While I have not performed this research yet, in the next chapter I will give some reasons to think this is true.

## 3.7. Conclusion: The challenge of the brain sciences

Let's conclude. The brain sciences have offered up the challenge that our sense of agency for our actions is an illusory phenomenon. What actually determines our actions is unconscious neural activity, which occurs well before, and apart, in the brain, from the neural activity that realises our sense of agency. In this chapter I have shown that while all this might be true for actions caused in service of our CAPs, there is no reason to suspect that it generalizes to the creation

(influencing) of those action procedures as well. This response makes use of a common response in the literature, namely that the instances of free action being studied are special in some way, but does so with a twist. I don't think these instances of free action being targeted by these studies are rare. In fact, I think they might reflect the majority of our free actions. If evidence of our free will is restricted to just a few actions and the creation (influencing) of our action procedures, then there is a natural tension with the prevalence of our sense of being free. While it's an empirical question how big a tension this is, it's plausible to think that our free action judgments were never just for a few special actions, such as the creation (influencing) of our action procedures, but always included the actions caused by those CAPs as well. From now on we will begin to roll out the machinery for indirect compatibilism, and in doing so begin to make sense of a different kind of freedom, indirect freedom.

## Chapter 4
## Indirect freedom

So far I have established that there is no reason, arising from the brain sciences, to think that our freedom-conferring mechanisms (henceforth FCMs) cause *none* of our actions. Unfortunately, it remains very plausible that FCMs cause very few our actions. If only a few actions get to count as free, then there will likely be a large discrepancy between the number of instances in the world we judge to be free, and the number of instances in the world where we are in fact free. I have called this the discrepancy problem (see §3.6.). Given this problem, if direct (conscious control) compatibilism is true of our concept of free action then, quite likely, it will turn out we don't have free will. As I noted earlier (see §3.6.), that's because it is plausible that we only have free will if enough of the cases we judge to be instances of free action are in fact free. If our sense of being free is as common as it seems to us to be, then according to direct compatibilism our sense of being free would be mistaken more often than not. In fact, the majority of the instances in the world we judge ourselves to have acted freely would instead be instances where our freedom was not exercised at all.

My proposed solution to the discrepancy puzzle takes the form of a new account of free action: indirect (conscious control) compatibilism. Recall that according to indirect compatibilism, we are not, as direct compatibilism tells us, only free when we perform an action that was directly caused by FCMs. We are also free when we perform an action that was indirectly influenced by FCMs. Put another way, an action can be free provided it can be traced back to an earlier instance of direct freedom. Unoriginally, I will call this latter kind of freedom, **indirect freedom**. While I will provide a precise definition of indirect freedom very shortly, here is a rough characterization to carry you over for now. An action is indirectly free if it's caused by an action procedure that was either created by the agent, could be destroyed by the agent, can be triggered on by the agent, can be monitored and turned off by the agent, or can be updated by the agent. Remember that an action procedure is a sub-personal level mechanism that causes actions, or influences other action procedures, both in a contained and reliable fashion.

A major upshot of my account is that when we report a sense of agency, even for an action as simple as a wrist flex, a left-or-right button press, or even basic addition-or-subtraction, and so on, we are correctly reporting an action performed with agency (as long as (as is typically the case) the action is the product of an action procedure which is influenced by an FCM). More than that, though, we are correct in thinking we perform such an action freely.

Here, then, is the plan for this chapter, divided into four sections. Firstly, in §4.1, I will provide a definition of indirect freedom, and describe in detail the conditions for an action to be indirectly free. Put loosely for the moment, an action is indirectly free provided it's caused by a **controlled action procedure**. A controlled action procedure (henceforth CAP) is an action procedure that was influenced by a FCM. For example, the act of flexing one's wrist during a Libet-style paradigm is indirectly free, if the action procedure which caused the wrist flex was created (influenced) by the agent. Then, in §4.2, I will continue addressing some evidence from the brain sciences that *prima facie* appears inconsistent with our having indirect freedom. Specifically, I will address in detail the contribution of various parietal regions, which I set aside in §3.5. After establishing that indirect freedom is not straightforwardly ruled out by current empirical evidence in the brain sciences, I will argue, in §4.3, that my new account of free action, **indirect compatibilism** is a better account of our concept of free action than is direct compatibilism. Finally, in §4.4, I will conclude.

### 4.1. Defining indirect freedom

Let's begin with my definition of indirect freedom. An action *A* is indirectly free, if and only if all of the following conditions are met:

1) [Controlled action procedure] *A* is caused by a *CAP*.
2) [No defeaters] There exist no defeaters to our free action.

Before continuing, let's also define what I mean by *CAP* in (1). I will define this in terms of five conditions I will call **direct control conditions**. It is sufficient for an action procedure to be a CAP if at least one of these conditions is met. It is also necessary for an action procedure to be a CAP that at least one of them is met. It's thus both necessary and sufficient that at least one of the direct control conditions is met for an action procedure to be a CAP, though of course more of the conditions might be met.

The *direct control conditions* are:

2a) [Creation] The action procedure was created by a directly free act of the agent.
2b) [Destruction] The action procedure could be destroyed by a directly free act of the agent.
2c) [Triggering] The action procedure was triggered on by a directly free act of the agent.
2d) [Monitoring] The action procedure is monitored by an FCM and could be turned off by a directly free act of the agent so chooses.
2e) [Update] The action procedure was updated by a directly free act of the agent.

I have called these conditions direct control conditions because they describe the various ways an agent can directly influence the function of an action procedure with an FCM, and so bring it

about that they cause indirectly free actions. In what follows I will unpack each of these direct control conditions and the definition of indirect freedom in much more detail.

### 4.1.1. Unpacking the definition of indirect freedom

The easiest way to unpack the definition of indirect freedom is by using an example. For ease and consistency, I will continue to use the example of participants' performance during Libet-style paradigms. However, it's important to realize that my definition of indirect freedom extends beyond the experimental context.

When someone successfully performs in a Libet-style paradigm, they repeatedly perform some simple action, or basic cognitive operation such as simple arithmetic, for an extended period of time. In Chapter 3, I established that while participants might consciously select amongst, cause, and guide, their actions over the course of the experiment, this is incredibly unlikely. Not only do the instructions in these experiments explicitly ask participants not to perform in this manner, participants' self-reports suggest they are complying with the instructions as best they can. Instead, participants might be performing Libet-style paradigms with a CAP: an action procedure created (influenced) by an FCM to unconsciously cause actions.

In the case of a Libet-style paradigm, participants might adopt a CAP in order to successfully perform the experiment with minimal cognitive effort. The created (influenced) action procedure would be a simple sub-personal level mechanism that would cause a simple wrist flex, a left-or-right button press, perform simple addition and subtraction, and so on, whenever the onset of an experimental trial was cued. Even if participants are unconscious of the process by which their actions are caused, simple actions such as flexing one's wrist can be indirectly free, provided they are caused by a CAP. Of course, for indirect freedom it's not enough for an action to be caused by any sub-personal level mechanism, the action procedure must have been influenced by an FCM. I have collectively called the different ways agent's can influence their action procedures with an FCM: direct control conditions. With this example in hand, let's begin working through the different direct control conditions and my definition of indirect freedom. An action is indirectly free, if and only if:

*4.1.1.1. [Controlled action procedure]: A is caused by a CAP.* Previously, I spelt out what it takes for an action procedure to be a CAP in terms of the following direct control conditions: [Creation], [Destruction], [Triggering], [Monitoring], and [Update], each of which I unpack shortly. However, before continuing I want to flag a few things. I will call the cases where all 5 direct control conditions are met **paradigmatic controlled action procedures**. Further, I will call cases where the subsets of direct control conditions are met **non-paradigmatic controlled action procedures**. This distinction will become important in Chapter 5 when discussing certain

non-paradigmatic controlled action procedures that might motivate arguments against indirect compatibilism, but much more on that later. For now, let's get clear on each of the different direct control conditions: [Creation], [Destruction], [Triggering], [Monitoring], and [Update]. An action procedure is a CAP, if:

*4.1.1.2. [Creation]: The action procedure was created by a directly free act of the agent.* Once again, let's return to the example of participants' performance during a Libet-style paradigm. There, a participant repeatedly performs a simple action (e.g., wrist-flex, left-or-right button press, simple addition or subtraction, and so on) in a timely manner, following some task cue, for the duration of the experiment. It's plausible to think that the simple action in question is being caused by a CAP, because the action procedure was one that was created by the participant. Upon being told the instructions for the Libet-style paradigm, the participant created, and implemented, an action procedure to unconsciously cause actions required by the task. That CAP makes it so that when a task cue is presented, the necessary actions are unconsciously caused in a timely manner. The actions caused by the CAP are indirectly free precisely because that action procedure that caused them was itself caused by a directly free action.

*4.1.1.3. [Destruction] The action procedure could be destroyed by a directly free act of the agent.* As well as creating action procedures, we should also expect that agents will be able to destroy their action procedures, should they freely want to do so. Destroying an action procedure, then, reflects an agent's capacity to erase an action procedure they possess. Having such a capacity would be valuable for whenever an action procedure is no longer required, or has become detrimental to the agent because of changes in the environment. For example, it's easy to imagine ourselves creating, and implementing, an action procedure that automatically causes a simple action that is consistently rewarded whenever it is cued in the environment. But imagine the environment changes so that the unconsciously caused action, which was consistently rewarded, is now consistently punished. It's now easy to imagine ourselves being motivated to destroy that CAP as soon as possible!

But how exactly do you destroy an action procedure? Unfortunately, this is not a matter I can settle in this thesis. Still, I think we must possess some mechanism that can allow us to destroy an action procedure in some way or another. While speculation about the nature of the mechanism or mechanisms that allow us to destroy action procedures is interesting (assuming they exist), I think my view is much more interesting if I am able to suggest some plausible mechanisms that might actually do this kind of work. As a result, before continuing on with my explication of indirect freedom, I will briefly suggest two mechanisms that might underpin our capacity to destroy action procedures, one *direct*, and one *indirect*.

*4.1.1.3.1. Destroying action procedures directly.* Let's start with a direct mechanism. It's possible we can destroy an action procedure simply by intentionally forgetting it. Such a capacity might reflect the same mechanism, or kind of mechanism, that has been implicated in cases of intentionally forgetting memories. There is suggestive, and accumulating, evidence that we have the capacity to intentionally forget learned word-pair associations (for example, see, Anderson, & Green, 2001; Fischer, Diekelmann, & Born, 2011; for review, see, Anderson, & Huddleston, 2012), and perhaps more impressively, certain autobiographical memories as well (Noreen, & MacLeod, 2013; Ryckman, Addis, Latham, & Lambert; 2018; Ryckman, Addis, Latham, & Lambert, manuscript). While the capacity to intentionally forget memories is undoubtedly controversial, much of this controversy is over the capacity to intentionally forget particularly traumatic or highly valenced memories.[37] It's perhaps less controversial to think some prosaic process might exist that allows for the targeted forgetting of certain memories, such as word-pair associations in laboratory tasks, and perhaps similarly, certain action procedures as well. However, just as some memories might be impervious to intentional forgetting, some action procedures might be too. In Chapter 5, I will explore non-paradigmatic controlled action procedures where destruction of the action procedure is not possible.

*4.1.1.3.2. Destroying action procedures indirectly.* It's possible that we don't, strictly speaking, destroy action procedures, rather we 'effectively' destroy them by modifying them in various ways. Although 'destruction' here is not being used in the same sense as with intentional forgetting, this fact shouldn't bother us, as terminating performance of the targeted action procedure is the goal in both cases. To illustrate, consider once again a simple action procedure created and implemented to perform a Libet-style paradigm. The CAP causes a simple action, or performs a simple cognitive operation, some small time after the presentation of a task cue. Instead of intentionally forgetting the action procedure, we might destroy the action procedure by altering it in ways that prevent it from performing. For example, we might remove the cue component so the action procedure no longer causes an action, or we might remove the response component, so the action procedure no longer has an action to cause. Alternatively, we might set a long time-delay between the cue and action production, such that there would be no reasonable expectation an action would be caused even when the action procedure is implemented. I will have a few more things to say about altering action procedures when I come to describe the direct control condition [Update] for [Controlled action procedure].

[37]For example, see Bulevich, Roediger, Balota, and Butler, (2006), Erdelyi (2006), Hayne, Garry, and Loftus (2006), and Kihlstrom (2002).

Now that I have finished describing two kinds of mechanisms that might underpin our capacity to destroy action procedures it's time to return to unpacking the direct control conditions for [Controlled action procedure] by looking at the next condition - [Triggering].

*4.1.1.4. [Triggering]: The action procedure was triggered on by a directly free act of the agent.* While perhaps not immediately apparent, I already touched on the direct control condition [Triggering] in my discussion of the condition [Creation]. When an agent is faced with an environment where they are able to automate some, or all, of their actions, they may create an action procedure to automate and control their actions. However, merely creating an action procedure is not enough. In addition to creating an action procedure, the agent must press also the action procedure into service. This further step is what the condition [Triggering] refers to - pressing an action procedure into service.

To illustrate, consider a participant who had previously created and implemented an action procedure capable of automating actions during a Libet-style paradigm. Now imagine they are bought back into the laboratory at a later date to perform the exact same experimental task. So long as the original action procedure has not been destroyed, they already possess a CAP capable of automating their performance the second-time through the task. As a result, the agent doesn't have to create another action procedure; instead they can simply choose to trigger the original action procedure on once again to automate and control their task-relevant actions. The actions caused by the CAP are indirectly free because the action procedure that caused them was triggered on to perform by a directly free act.[38]

*4.1.1.5. [Monitoring]: The action procedure is monitored by the FCM and could be turned off by a directly free act of the agent so chooses.* When an agent is faced with an environment where they are able to automate some, or all, of their actions, they might create and trigger on an action procedure. However, sometimes we can find ourselves performing the wrong kinds of actions, or simply wanting to change the kinds of actions we perform and so want to turn off the current action procedure. This latter step is what the condition [Monitoring] refers to – the turning off an action procedure if we so choose.

To illustrate, consider a participant who during a Libet-style paradigm develops an itch. Imagine possessing an action procedure that functions by relieving itches as they occur on your body. Now let's imagine you develop an itch and the action procedure automatically causes a scratching action to occur at the site of the itch. It's not hard to imagine that we possess such an

[38]Strictly speaking the action caused in this example was also indirectly free because it was caused by an action procedure freely created by the agent. In Chapter 5 I will discuss cases where the conditions [Creation], [Destruction], [Monitoring], and [Update] are restricted or absent. There, it will become apparent that [Triggering] alone is sufficient for an act to count as indirectly free.

action procedure, nor is it hard to imagine that we possess such an action procedure despite never actually creating it ourselves. Perhaps some of us did in the past create, trigger on, or update an action procedure for scratching, through a directly free action, but let's imagine you didn't do any of these. Perhaps you might be able to destroy the action procedure, but again let's imagine you can't and that the scratching action procedure is effectively hardwired into us.

Despite all this, when you performed the action of scratching yourself you acted freely; after all it still seems that had you wanted to, you could have intervened and prevented yourself from scratching at the site of the itch by turning off the scratching action procedure. When an agent is faced with an environment where some, or all, of their actions are being caused by an action procedure, they may monitor their action procedures and, if they so choose, intercede to turn off action procedures that are producing the wrong actions, or, if they want, to perform actions of a different kind than what they are currently performing. Having now described [Monitoring], I can move onto [Update], the final direct control condition for [Controlled action procedure].

*4.1.1.6. [Update]: The action procedure was updated by a directly free act of the agent.* I touched on the direct control condition [Update] while previously discussing the condition [Destruction]. There, I suggested that one way an agent might destroy an action procedure they possess is by altering it so it no longer performs. However, just as an agent should be able to update an action procedure to prevent its performance so too should they be able to update an action procedure to change or improve its performance.

To further illustrate the direct control condition [Update], once again imagine a participant who possesses a CAP capable of performing in a Libet-style paradigm. Now imagine the participant is bought back to the laboratory to perform a slightly modified version of the task. While they are still required to perform some action, sometime after a cue is presented their body is now hooked up to an accelerometer. Each time they perform an action, feedback is presented about the speed of their action. The feedback indicates whether or not the action was performed at the correct speed, and when it indicates the action was performed at the incorrect speed, whether they should speed-up or slow-down their action. Instead of creating a new action procedure, the participant in this example might simply trigger on the original action procedure and update it, so the caused action comes to be performed at the required speed. Of course, the process of updating an action procedure might go through numerous iterations as the agent adjusts their action speed, and then looks to the feedback in order to see whether their update was successful. Still, once the correct speed has been identified, the updated action procedure will be able to automate performance in the exact same manner as the original action procedure.

The action caused by the updated portion of the action procedure is indirectly free because the updates to the action procedure were brought about by directly free actions.[39] With my description of the direct control condition [Update] complete, I have now finished describing the direct control conditions for [Controlled action procedure]. This leaves just the final condition for indirect freedom to describe: [Defeater].

*4.1.1.7. [Defeater] There exist no defeaters to our free action.* While there are numerous positive conditions for free action, there are also numerous negative conditions that, for the most part, we agree count against us acting freely. For example, most people think agents subject to physical restraint cannot act freely. On the other hand, as mentioned in the introduction, some people will think that agents who lack autonomy cannot act freely either. Whatever conditions stand as defeaters to acting freely will also stand as defeaters to acting indirectly freely.

**4.1.2. Conclusion: Defining indirect freedom.**

In summary, an action is indirectly free, if and only if, (1) it is caused by a CAP, and (2) there exist no defeaters to our free action. In order for an action procedure to be a CAP, 1 of 5 direct control conditions must be satisfied. Direct control conditions describe the various ways an agent can directly influence the function of an action procedure with an FCM, and so bring it about that they cause indirectly free actions. Either (a) the action procedure is one that was created by a directly free act of the agent or (b) the action procedure could be destroyed by a directly free act of the agent or, (c) the action procedure is triggered on to perform by directly free act of the agent or (d) the action procedure is monitored and could be turned off by a directly free act of the agent, if they so choose or, (e) the action procedure has been updated by directly free act of the agent. I have called cases in the world where all 5 direct control conditions are met – **paradigmatic controlled action procedures.**

According to indirect compatibilism, an agent not only acts freely in a few special cases (such as creating, destroying, triggering, monitoring, and updating action procedures), they also act freely when they perform an action caused by a CAP. For this reason, indirect compatibilism judges many more actions to be free than direct compatibilism. I think this result is correct, because it's faithful to our sense that when we act, we do so freely more often than not. With my explication of indirect freedom now complete, I will now (§4.2) make a detour back through brain sciences and describe some evidence that is *prima facie* inconsistent with us having indirect freedom. Primarily my focus will be on describing the contribution of neural activity observed in

[39]Strictly speaking the action caused in this example was also indirectly free because it was caused by an action procedure freely triggered by the agent. In Chapter 5 I will discuss cases where the condition [Creation], [Destruction], [Triggering], [Monitoring] are restricted or absent. There it will become apparent that [Update] alone is sufficient for an act to count as indirectly free.

parietal regions to Libet-style paradigms, findings I set aside in Chapter 3. In §4.3, I will argue that indirect compatibilism is a better account of our concept of free action than is direct compatibilism.

## 4.2. The challenge of the brain sciences: Parietal redux

To this point I have done two things. I've proposed a definition of indirect freedom, and described each of its conditions in detail. Before I can link indirect freedom (and indirect compatibilism for that matter) to the brain sciences, I need to take a step back and finish discussing some neuroscientific findings set aside in Chapter 3. These are findings which appear to show that neural activity for an action occurs well before the neural activity for FCMs. If this is true, then FCMs cannot have been what caused the action in the first place, and, as a result, the action is not free. Remember that FCMs are those conscious psychological processes (along with their neural correlates) that cause free actions, and hence are crucial for indirect freedom as well. In this section I will discuss these findings and offer up an alternative explanation for these results that is consistent with indirect freedom.

The findings I will be discussing concern the role of neural activity in parietal brain regions, specifically the precuneus and posterior cingulate, in the performance of Libet-style paradigms. Soon and colleagues (2008; 2013) suggest that activity in the precuneus could reflect the unconscious storage of a chosen action for future performance. Meanwhile, the posterior cingulate in conjunction with the other predictive regions might reflect activation of the brain's resting network, and provide a further reason to believe participants are unconscious of their task-relevant neural activity. After describing each of these suggestions, I will propose an alternative explanation of the data that is consistent with participants deploying FCMs, just as I did with prefrontal neural activity in Chapter 3. Not only is it plausible that action procedures might be triggered on by FCMs; parietal activity, *inter alia,* might also contribute to creating action procedures (and select actions), monitoring, and updating them as well.

### 4.2.1. Parietal neural activity and free action

The empirical data that I will be reinterpreting in this section once again comes from Libet-style experiments, so it will be helpful, before starting, to briefly describe the methodology and their major findings again. In perhaps the simplest Libet-style paradigm performed with concurrent fMRI, participants pressed a button with either their left or right hand whenever they first felt the urge to do so (Soon et al., 2008; though see also Bode et al., 2011; Soon et al., 2013). At the same time, participants fixated on a stream of letters presented in the centre of a screen. Once the participant had pressed either button, they selected the letter that was present on the screen

when they first felt the urge to press the button. Researchers found patterns of neural activity in certain brain regions predicted, above chance, the actions participants would perform. This predictive neural activity began up to 10 seconds before participants first reported their sense of agency for the action they performed. As a result, the researchers concluded that FCMs could not be causing our actions, as the onset of neural activity for the action occurred well before any reported sense of agency for that action. The brain regions identified as being the earliest predictors of actions were prefrontal brain regions of the frontopolar cortex and medial prefrontal cortex, and parietal brain regions of the precuneus and posterior cingulate. Here our focus will be on the possible roles of the precuneus and posterior cingulate.

In what follows, I will first discuss the role suggested by the experimenters. After that, I will then go on to describe some of the other possible roles that this neural activity might play. Once again, as with my previous discussion in Chapter 3, it should be noted there is no consensus yet regarding what common functional role parietal brain regions play in the performance of Libet-style tasks.

*4.2.1.1. The precuneus.* Let's start with the precuneus. Soon et al. (2008; 2013) suggest that the precuneus might store the unconsciously chosen action before it is triggered. In addition to their primary study, Soon, et al. (2008) ran a control task to distinguish the contributions of prefrontal neural activity and parietal neural activity. Rather than have participants perform their chosen action whenever they wanted, participants were asked to perform their chosen action only when cued by the experimenter. They found that while neural activity in the frontopolar cortex was predictive while selecting which action to perform, the precuneus was predictive while waiting for the experimenter to cue action performance. This result suggests that the precuneus is holding onto the chosen action prior to it being performed. Further, following from other literature, once performance has been cued, the greater posterior parietal cortex is capable of causing and guiding the action (for example, see Connolly, Anderson, & Goodale, 2003; Tosoni, Galati, Romani, & Corbetta, 2008; Karl, & Whishaw, 2013).

*4.2.1.2. The posterior cingulate.* Now onto the posterior cingulate, Soon, et al. (2013) observed a significant overlap between the posterior cingulate and other predictive brain regions, and the brain's 'default mode network'. The default mode network is a task network in the brain that appears active whenever there is no demanding, externally directed task being performed (e.g., a difficult perceptual task). Colloquially, this network has been associated with 'mind wandering', and related cognitive abilities such as self-referential thought, remembering the past, simulating the future, and perspective taking (for example, see Raichle, MacLeod, Snyder, Powers, Gusnard, & Shulman, 2001; Greicius, Krasnow, Reiss, & Menon, 2003; for review, see Snyder, & Raichle,

2012). Soon, et al. noted that the levels of neural activity in the default mode network in the early stages of an experimental trial were the same as those observed during rest-periods in other experimental tasks. If participants are mind wandering, then they might not be thinking about experimental performance. As a result, we have another reason to think that participants are not conscious of the predictive neural activity for task performance in Libet-style tasks. *4.2.1.3. My initial assessment.* Consistent with my discussion in Chapter 3, I think these researchers are generally right regarding the functional role of these brain regions in the vast majority of experimental trials. However, the fact that participants are unconscious of the production of their actions doesn't mean they are not free. Rather, participants are unconscious of the production of their actions in Libet-style paradigms because they are being caused by a CAP. This means the actions are indirectly free and is evidence for our free will according to indirect compatibilism. What I am interested in here is getting clear on what this unconscious neural activity might reflect, and whether it's plausible, on occasion, to think it might reflect FCMs causing actions. After all, for an action to be indirectly free, it's not enough that it be caused by an action procedure. The action procedure must be one that was created, could be destroyed, triggered, monitored, or updated by the agent.

**4.2.2. Reinterpreting the role of the precuneus**

Here is my proposal of what neural activity in the precuneus might reflect in a Libet-style task. Neural activity here might be, on occasion, a candidate realiser of FCMs performing a role in causing an action, or in influencing an action procedure. For readers that are interested, §4.2.2.1 contains some empirical evidence to support my proposal. Instead of realising the unconscious storage and cause of an action, neural activity in the precuneus might realise a participant holding *some* actions in mind, simulating their performance, and then choosing one of them to be caused. As long as you think it's plausible that more than one action might be stored prior to performance, the predictive neural activity observed in these studies might simply realise the agent's own preferences or biases towards a particular action for any given experimental trial. After all, there is nothing about this process that requires all stored actions be equally favoured to be chosen. Furthermore, even if this occurs unconsciously, it gives us a reason to doubt the earliest signs of predictive activity reflect some chosen action.

Of course, I don't think this alternative account is representative of performance during a typical experimental trial. After all, it makes little sense to commit so many cognitive resources to continuously hold in mind, simulate, and cause, individual actions, especially when the performance is not incentivized in any manner. Still, this alternative account might capture performance at the beginning of the experiment when the participant is still creating an action

procedure to automate performance during the task, or if they already possess a suitable action procedure, updating and triggering it on back into service.

*4.2.2.1. Reinterpreting the role of the precuneus: evidence*. In what follows, I will describe some empirical evidence in support of my proposal regarding the role the precuneus might play in Libet-style tasks. If I am right, then there's a reading of the results of these experiments that is consistent with the participants in these experiments continuously acting in a directly free manner. More plausibly, though, there is a reading that is consistent with the participants in these experiments continuously acting in an indirectly free manner; provided they performed at least one directly free action, either to create, or trigger on, an action procedure to automate their performance in the task.

I have divided my presentation of the evidence into two parts. First, in §4.2.2.1.1 I will present evidence in support of my claim that neural activity in the precuneus might at times realise the storage of a number of different possible actions. This is important as it means the predictive neural activity observed in the precuneus in Libet-style tasks might not reflect a single action, but instead the agent's own preferences or biases towards a particular action in an experimental trial.

Then in §4.2.2.1.2 I will present evidence to support my assertion that neural activity in the precuneus might at times be a candidate realiser of an FCM. For example, as well as simply storing a number of different possible actions, neural activity in the precuneus might also realise the agent consciously holding these different possible actions in mind and simulating them. This is important as it means that the neural activity in the precuneus in Libet-style tasks might, on occasion, play a role in causing actions which are *directly* free. This includes the act of creating, destroying, triggering, monitoring and updating actions procedures that automate participants' performance. Any action procedures which are influenced by these directly free actions will in turn cause indirectly free actions. This is true even if these researchers are correct about participants being reliably unconscious of the processes that lead to the actions being caused in these tasks.

*4.2.2.1.1. Evidence for the precuneus storing multiple actions.* Let's begin with some evidence that neural activity in the precuneus might not just be for the unconscious storage of a single action. Assume for the moment that neural activity in the precuneus does reflect the unconscious storage and cause of actions. I see no reason to think it reflects the storage of a *single* chosen action, rather than some number of possible actions that could be used to perform the task. For example, in the Libet-style paradigm described earlier, why think that it's *just* the action to press the button with your left hand, or press the button with your right hand, that is unconsciously

stored up to 10 s prior to actual performance. Both simple actions could easily be stored. In addition, storing both actions secures some level of behavioural flexibility, especially if one of the actions turns out to be unviable prior to being caused.

Evidence from both behavioural and transcranial magnetic stimulation studies suggest that multiple motor actions might be stored and caused from the precuneus (for example, see Oliveira, Diedrichsen, Verstynen, Duque, & Ivry, 2010; though see also Johnson, 2000). In speeded reaching tasks participants perform the worst when the 'reach target' is placed equidistant from both hands. This might appear strange, but it makes sense if the reaching action for both hands is being stored, ready to be performed. As neither the left nor right hand is favoured, a conflict occurs in action selection that must be resolved in order for an action to be performed. Second, the application of transcranial magnetic stimulation over the surface of the left posterior parietal cortex results in participants making many more reaching responses with the left hand. As a reminder, motor control is contralateral in the brain, which means the left hemisphere stores and guides right handed actions. Presumably this result reflects the active disruption of the right hand action in some way. If the predictive neural activity present in the precuneus doesn't reflect the unconscious storage of a single action, it's plausible that it might instead reflect the agent's own preferences or biases towards a particular action in an experimental trial.

*4.2.2.1.2. Evidence of the precuneus as a candidate realiser of FCMs.* Now, let's move onto some evidence that neural activity in the precuneus might be a candidate realiser of an FCM. Important for my purposes is evidence that suggests that neural activity in the precuneus might contribute to actions which are directly free. This includes the act of creation, destruction, triggering, monitoring, and updating action procedures that automate participants' performance. Remember that if an action procedure is created, could be destroyed, triggered, monitored, or updated by a directly free act of the agent, then the actions it causes are indirectly free. This is true even if these researchers are correct about participants being reliably unconscious of the processes that bring about their performance on these tasks.

Once more, I see no reason to think that neural activity observed in the precuneus must realise the unconscious storage and cause of an action. Due to its extensive projections to other cortical and subcortical regions the precuneus serves as a vital hub integrating our own internal thoughts with information coming from the environment (for review, see Cavanna, & Trimble, 2006).[40] Cognitive capacities relevant for free action associated with neural activity in the

[40]Other important functional roles for the precuneus include autobiographical memory (for example, see Fletcher, Frith, Baker, Shallice, Frackowiak, & Dolan, 1995; Gilboa, Winocur, Grady, Hevenor, & Moscovitch, 2004), music

precuneus include: endogenous control of visual attention (for example, see, Culham, Brandt, Cavanagh, Kanwisher, Dale, & Tootell, 1998; Simon, Mangin, Cohen, Bihan, Dehaene, 2002; Nagahama et al., 1999), and mental imagery during motor simulation and mental navigation (for example, see Ghaem et al. 1997; Malouin, Richards, Jackson, Dumas, & Doyon, 2003; Ogiso, Kobayashi, Sugishita, 2000), deductive reasoning (Knauff, Fangmeier, Ruff, & Johnson-Laird, 2003), empathy and forgiveness (Farrow et al., 2001; Ochsner et al., 2004). Further to this, activity in the precuneus has been associated with self-referential processing (Kircher et al., 2000; Kircher, Brammer, Bullmore, Simmons, Bartels, & David, 2002), including self-awareness (Kjaer, Nowak, Kjaer, Lou, & Lou, 2001), and perhaps most importantly our experience of agency (Farrer, & Frith, 2002). Importantly then, neural activity in the precuneus appears to be closely associated with our capacity to simulate, plan and reason: paradigm examples of FCMs.

**4.2.3. Reinterpreting the role of the posterior cingulate**

So far I have discussed the role of the precuneus in Libet-style paradigms. More often than not neural activity in the precuneus might realise the unconscious storage and cause of participants' actions. Importantly though, it might on occasion be a candidate realiser of an FCM, such as holding in mind and simulating various actions. I will now turn to discuss one final finding that suggests participants are unconscious of task related neural activity during Libet-style paradigms. Recall, Soon et al. (2013) noted that levels of neural activity in the predictive brain regions, including the posterior cingulate cortex, overlaps with a task network in the brain known as the default mode network (for review, see Snyder, & Raichle, 2012).[41] Further, they noted that neural activity in the default mode network at the time the earliest predictive activity was observed was what you would expect if the participant was at rest (i.e., not performing the experimental task). As I described earlier in §4.2.1.2., the default mode network is commonly associated with participant's minds wandering; varieties of self-referential thought, including remembering the past, simulating the future, and perspective taking. *Prima facie*, this result rules out the alternative account I suggested previously in §4.2.2. If neural activity in default mode network reflects participant's minds wandering, then we have some reason to think they are unconscious of any task-relevant neural activity.

I will make two proposals in response. Once again these proposals are meant to show that this evidence is compatible with FCMs causing actions, or perhaps, more likely, influencing

---

perception (Satoh, Takeda, Nagata, Hatazawa, & Kuzuhara, 2001), and mental rotation (Suchan et al., 2002). There are still more examples that might be listed.

[41]I will mainly restrict my discussion on the posterior cingulate cortex to networks of brain regions. This is not a reflection on its importance. The posterior cingulate appears to be associated with directing attention either inward on one's own thoughts, or outward into the environment. It also appears to be important for the transmission of information across different brain networks (for review, see, Leech, & Sharp, 2014).

action procedures. Interested readers can find some empirical evidence in support of my proposal in §4.2.3.1. First, it matters what people are doing when they are mind wandering. Sure people might not be conscious of task-relevant neural activity for the experimental task, but if experimental performance is automated they might be conscious of some other task-relevant neural activity. For example, planning what they want to do when the experiment is over, spend their participation payment on, cook for dinner, and so on. Put another way, while these results might not be evidence of directly free actions in the experiment, this might have been the wrong place to look. The real test is whether the neural activity people are conscious of is causally relevant for actions caused after the experiment.

Second, according to indirect compatibilism it's not necessary for free action that the agent be conscious of the neural activity that caused it. Every action in the experiment can be free, provided the action procedure that caused them was created, could be destroyed, triggered, monitored, or updated by the agent. If you think it's plausible the activity observed here might reflect, *inter alia,* planning future actions, then it's plausible it might reflect the creation, destruction, triggering, monitoring, or updating of an action procedure to control and automate performance in the experiment. This might have occurred in a very early experimental trial, or perhaps even before the experiment commenced.

*4.2.3.1. Evidence of the posterior cingulate as a candidate realiser of FCMs.* I think the kinds of capacities associated with neural activity in the default mode network make it a candidate realiser of some FCMs. These capacities include self-referential thought, remembering the past, simulating the future and perspective taking. Important for my purposes is evidence that suggests that neural activity in the default mode network contributes to actions which are directly free. Recall, this includes the act of creating, destroying, triggering, monitoring, and updating action procedures that automate the participants' performance. Provided that an action procedure is created, could be destroyed, triggered, monitored, or updated by a directly free action of the agent, then the action it causes will be indirectly free. This is true even if participants are unconsciously performing actions for the majority of the time they spend on these tasks.

If I am right that the default mode network contributes to actions which are directly free, then once again I think there's a possible reading of these experimental results that is consistent with the participants in these experiments continuously acting in a directly free manner. More plausibly though, the participants in these experiments continuously act in an indirectly free manner provided they performed at least one directly free action, either to create, trigger, or update an action procedure to automate their performance in the task.

I see no reason to think that neural activity in the default mode network does not contribute to directly free actions, or that participants were not conscious of task-relevant activity at any stage in these experiments. First, there is a significant overlap between the default mode network and the fronto-parietal control network, a network we observe to be active when people are engaged in cognitive control (for example, see Vincent, Kahn, Snyder, Raichle, & Buckner, 2008; Margulies et al, 2009; Leech, Kamourieh, Beckmann, & Sharp, 2011). Second, there's not just a significant overlap between the default mode network and cognitive control network, but significant interaction between them as well (for example, see Smallwood, Brown, Baird, & Schooler, 2012; Cocchi, Zalesky, Fornito, & Mattingley, 2013; Beaty, Benedek, Kaufman, & Silvia, 2015). While the default mode network holds various internal trains of thought in mind, the cognitive control network is able act upon them in a targeted manner (e.g. Smallwood, et al., 2012). Importantly, for my purposes, if that's right then participants are not just conscious of this neural activity, this neural activity is a candidate realiser of FCMs.

For example, perhaps a participant wonders what they want to do after the experiment and so recalls some activities they have done in the past. With these in mind said participant goes about choosing the one they desire the most, or else have the most powerful reasons to favour. However, maybe they also want to include their partner in their plans, so simulate and hold in mind their partner's preferences while preparing to make their choice. One thing a participant can do is create, trigger, or update, an action procedure to automate experimental task performance while they consciously deliberate over more important matters. If that's right, then it's plausible that this neural activity could sometimes reflect the creation, triggering, or updating an action procedure for simple experimental task performance as well.

#### 4.2.4. Action procedure [Update] and the brain sciences

Before concluding, I want to make a very brief note about error detection, unconscious learning, and, most importantly, consciously updating our action procedures. One important direct control condition is the capacity to alter and update the function of an action procedure with an FCM. Unfortunately, there's no evidence in Libet-style paradigms that speaks directly to this condition. That's because performance in these kinds of experiments is never incentivized in any manner, nor are there any expectations about performance, aside from that some action is caused in a timely manner. Much later, in §5.5 and Appendix A, I will return to provide some evidence from my own investigations for this condition. For now, though, we can imagine slight variations of these Libet-style paradigms where performance is incentivised in some manner, or is classed as either right or wrong. For example, earlier I described using an accelerometer so that an action might be performed at a correct or incorrect speed. But there are numerous other

possibilities you could imagine that would have the same effect. Here, I will restrict myself to outlining two brain regions that are important for updating action procedures, whose activity is commonly indexed by the electrocortical components: the error-related negativity (erN) and error related positivity (erP).

*4.2.4.1. Error-related negativity.* Firstly, the erN originates in the anterior cingulate cortex and appears when people have made an error (Holroyd & Coles, 2002; Gehring, Goss, Coles, & Meyer, 1993). Importantly, the magnitude of the erN is not tied to the conscious awareness of the error. Rather, it's tied to the degree to which a participant learns after an error has occurred (Frank, Woroch, &, Curran, 2005). This means it's possible for someone to make a mistake, adjust for their performance and carry on, without ever being aware of the error or the adjustment. For the direct compatibilist these unconscious alterations would not be free as they are not the result of FCMs. On the other hand, for the indirect compatibilist, provided the action procedure that caused the action was influenced by an FCM, these unconscious updates would be judged to have occurred freely.

*4.2.4.2. Error-related positivity.* The erP is thought to originate in the posterior cingulate cortex, which I have discussed already in the context of the default mode network and cognitive control network (Vocat, Poutois, & Vuilleumier, 2008). Unlike the erN, the erP does appear to reflect conscious awareness an error has occurred (Ullsperger, Harsay, Wessel, & Ridderinkhof, 2010). Interestingly, while the magnitude of the erP appears to reflect the strength of evidence that an error has occurred, the strength of this evidence does not guarantee that someone will make any changes to their performance (Steinhauser, & Yeung, 2010; though see also Orr, & Carrasco, 2011). This might appear strange at first. But, when deciding what to do when an action procedure makes errors, the evidence that errors are occurring might be only one of the things we take into consideration. For example, imagine a speeded-response task where you are required to make as quick a response as possible when cued, with either your left or right hand. Due to the task requirement to respond as quickly as possible, you might acquire plenty of evidence for errors (i.e., responding too early, with the wrong hand, and so on). Still, you might refrain from updating the action procedure and persist with it in its current form, as it's plausible that improving performance accuracy could result in you responding slower than you are able to. The evidence reflected by the erP might act as a source of evidence for some FCM to update an action procedure. Alternatively, in some cases it may provide the impetus to create an entirely new action procedure.

**4.2.5. Conclusion: The challenge of the brain sciences: Parietal redux**

In summary, according to Soon, et al. (2008; 2013) the neural activity we observe in the precuneus might reflect the participant unconsciously storing and causing an action. Further, the network of predictive brain regions, including the posterior cingulate cortex, might reflect the participant's mind wandering, unaware of any task-relevant neural activity. In this section I have shown that it's plausible that this neural activity might also reflect participants' FCMs causing actions or influencing action procedures. These results are consistent with their being at least *some* free actions according to direct compatibilism and *numerous* free actions according to indirect compatibilism. In the next section, I will argue that we should favour indirect compatibilism over direct compatibilism, not just as an account of this empirical evidence, but of our concept of free action as well.

### 4.3. Why indirect compatibilism?

Over the course of my discussion of some of the most important empirical studies into free will (e.g., Libet et al., 1983; 1985; Bode et al. 2011; Fried et al., 2011; Soon et al. 2008; 2013) I established two things. First, current evidence from the brain sciences fails to show there are *no* free actions according to direct compatibilism. Second, that same evidence does appear to show that there might be far fewer free actions, according to direct compatibilism than we might have supposed given the number of cases we judge to be free. Thus, while the evidence might fail to show there are *no* free actions, it might still show that there is *no* free will. That's because, as I have repeatedly said, free will as a general capacity requires more than there being some free actions; it requires that we act freely in a reliable fashion.

Of course, I have continually suggested there are possible readings of the current evidence that are consistent with participants' FCMs causing their actions in these experiments. Participants might have, *inter alia,* consciously deliberated, chosen an action, observed the consequences of their action and then consciously adjusted their performance accordingly. However, despite such a reading being possible, we have good reasons to believe people are not performing these experiments in this manner. Thus, while it's plausible there might be some free actions of the kind suggested by direct compatibilism, there are likely to be far fewer than the number of cases in the world we judge to be free.

In this section I argue that my new account of free action, indirect compatibilism, not only makes better sense of our judgments regarding cases of free action, but we have good reason to prefer it to direct compatibilism. The structure of this section will follow the structure of an argument for this conclusion.

#### 4.3.1. One argument for indirect compatibilism

Here is my argument for indirect compatibilism:

Argument for indirect compatibilism

1) [*Plausible* empirical claim] Indirect compatibilism vindicates as many, or more, of our pre-theoretic judgments regarding which actions are free (and which actions are unfree) than does direct compatibilism.
2) [Principle of theory selection] If two accounts are equally virtuous in other respects, but one makes judgments better aligned with out pre-theoretic judgments, we should prefer the account that is better aligned with our pre-theoretic judgments.
3) [Theoretical virtues] Indirect compatibilism and direct compatibilism are equally virtuous in other respects.
4) [Conclusion] Therefore, indirect compatibilism is a better account of our concept of free action than is direct compatibilism (from 1, 2, and 3).

The argument for indirect compatibilism is valid. The majority of my discussion here will be dedicated to motivating Premise (1). That is, I will establish there are excellent reasons to think that indirect compatibilism vindicates more of our pre-theoretic judgments regarding which actions are free than does direct compatibilism. Later on, I will say a little bit to motivate Premise (2) and (3), but I think it should be relatively uncontroversial that it's theoretically virtuous for an account of free action to align with our pre-theoretic judgements on cases of free action. Similarly, provided rival theories are equally virtuous with respect to other theoretical considerations, we are warranted in choosing the theory that best aligns with our pre-theoretic judgments.[42]

*4.3.1.1. Motivating Premise (1).* Here's my plan to motivate Premise (1). First, it's important to remember what we want our account of free action to do with respect to our judgments on cases of free action. We are seeking an account that can make sense of as many of judgments on cases of free action as possible. By articulating a broader set of conditions under which actions are free (including both directly and indirectly free actions), indirect compatibilism is able to account for many more of our free action judgments than does direct compatibilism. Of course, there is a theoretical cost that must be paid for the inclusion of these further conditions, but, as I will show, we have reasons to think that indirectly free actions might already be part of the folk concept.

Second, I will address some concerns about indirect compatibilism as a viable account of our concept of free action. It's not enough that an account of free action makes sense of our free

[42]Some common theoretical virtues include empirical accuracy, causal adequacy, explanatory depth, internal consistency and coherence, universal coherence and unification, fruitfulness, parsimony, and applicability (for recent discussion see Keas, 2018).

action judgments, the account itself must also be one that is deemed appropriate by the folk. At this time, I don't possess the data that shows the folk accept the conditions associated with indirect freedom as part of our concept of free action. It's possible that upon further empirical testing I discover that people reject the additional conditions required for an action to be indirectly free. *Prima facie* this would indicate that indirect compatibilism is not as good as direct compatibilism conceptually for the folk. However, in that case I will argue that it is nearly as good conceptually, despite the cost that would be incurred to make principled revisions to our concept of free action. That's because if we can't successfully account for those cases that would have been picked out as indirectly free, we not only risk discarding most of our free actions, but the fact that we have free will at all.

*4.3.1.2. Indirect freedom as part of the folk concept?* In Chapter 2 of this thesis I used people's judgments for cases of free will as datum that any prospective account of free will is required to explain. Unfortunately, it's not enough to simply uncover our 'raw' judgments and construct an account from them. Our raw judgements are liable to be noisy, conflicting and even false. In addition, in some cases there might be a lack of consensus for the correct judgment, and some cases might not be considered as important as others to the prospective account. For these reasons, before an account of free action can be constructed, people's raw judgments for cases must be systematized in some manner. In Chapter 2 of the thesis I subjected people's raw judgments to empirical scrutiny; but that's not the only kind of methodology we might employ (though it is my preference). We can also idealize from these raw judgments in order to identify which judgements need be explained by the prospective account, which judgments need to be discarded, which judgments reflect closely related concepts that happen to share the same reference term, and so on. As I have not yet performed the necessary empirical studies required to test whether indirect freedom is satisfactory for the folk, for now I'll be idealizing from some of our raw judgments on cases of free action. If you find this problematic, treat what I have to say as a series of plausible hypotheses about what I expect to observe once the necessary empirical studies have been performed.

The simplest reason I think indirect compatibilism is better, conceptually, than direct compatibilism is because it gets many more of our free action judgments correct. Throughout this part of the thesis, I have used the case of actions performed in Libet-style paradigms, for good reason. It's highly plausible that these actions are not free according to direct compatibilism, and are free according to indirect compatibilism. It might be countered that no one ever really believed that these kinds of simple actions were the kind of actions that we considered to be free to begin with. However, I think this is incorrect. Scientists looking to

investigate free will must have thought that these simple actions were the sort of actions that we would judge to be free. Further, there is at least some evidence that what guided them to these actions was reflection on the folk concept of free action. While not discussed much in the empirical literature, Libet (1985) offers some justification of why we should consider the actions in these experiments as free. The most important reason he writes is [*emphasis* his own]:

> "...subjects *feel* introspectively that they are performing the act on their own initiative and that they are *free* to start or not to start the act as they wish." (pg. 529)

For the most part, I assume the same rationale (or something sufficiently similar) permeates through the brain sciences and other empirical investigations into free will. The reason these kinds of actions are relevant for free will is because they are kinds of actions that we judge, upon reflection, to be free.

Many brain scientists have claimed their empirical research shows there is no free will because the candidate realisers of FCMs could not be what are causing our actions in these experiments. However, there is an assumption driving this conclusion which I have been at pains to suggest is mistaken. The conclusion that there is no free will tacitly assumes that the reason we judged these cases to be free was because they were cases of directly free actions. In the cognitive sciences our conclusions are directly informed by our experimental task designs. If I'm interested in a specific cognitive capacity, then I will design an experiment so that only the cognitive capacity of interest is impacted (or perhaps differentially impacted) by the task design. Of course I might not succeed in my design. Perhaps a number of different capacities are impacted, or there are individual differences in the way people perform the task. Each of these factors can give rise to a number of confounding explanations for the observed result. Assuming for the moment that participants did perform Libet-style paradigms by repeatedly performing directly free actions, then the evidence might suggest we don't act freely in these cases, and thus don't have free will. But there is no reason to think participants perform these experiments in this manner, nor is any reason to suggest people think they perform these experiments in this fashion either.

*4.3.1.3. Indirect freedom and principled revisions to the folk concept.* Indirect compatibilism is entirely consistent with our judgment that some simple actions, such as those performed in Libet-style paradigms, are free. Unfortunately, at this time there's an open empirical question whether the sufficient conditions introduced to pick out indirectly free actions are acceptable to the folk. I still need to test whether the folk judge indirectly free actions to be free, or whether they would accept indirect freedom in order to secure certain cases in the world as being free.

At the end of Chapter 3 (§3.6) I described my approach to the construction of the folk theory as a combination of theory-based and case-based approaches. This means the construction of the folk theory occurs alongside the systematization of people's judgments. To some degree, this combined approach shares some key similarities to Rawls' (1971) influential methodology of reflective equilibrium, as we go back and forth between our prospective theory and our judgments on cases, until equilibrium is reached. In my case, the relevant back and forth concerns a prospective account of the concept of free action and our judgments for cases of free action in the world. Still, you might wonder why a notion such as indirectly free action, or something sufficiently similar, would make it into the final theory. One reason I have continued to draw upon comes from simple reflection upon one's experience as an agent: we feel as though when we act we do so freely more often than not. If you are tempted by that thought, then direct compatibilism is entirely incapable of accommodating the number of cases we judge to be free. Indirect compatibilism can supply many more cases in a principled manner, as FCMs cause actions that are both directly and indirectly free.

There is a possibility that upon empirical testing I might discover that the folk reject the conditions required for indirectly free actions, and as a result reject my new account of free action, indirect compatibilism. Instead they might respond in a manner that is consistent with direct compatibilism being the correct account of the concept of free action. I think this result is pretty unlikely, as there is a natural tension between the costs associated with revising the folk theory to include the conditions required to judge actions as indirectly free, and having to discard the majority of cases in the world that we judge to be free. Further, I think this tension tips us towards accepting the costs associated with revising the folk theory of free action when we consider the consequences of discarding the majority of purported cases of free action in the world. As I have repeatedly noted, I think our general capacity for free will requires that we act freely in a reliable fashion. Discarding the vast majority of cases we judge to be free is tantamount to us accepting we don't have free will. Faced with such a decision, I think we would revise our concept of free action, as the alternative is to accept there might be some free actions, but no free will. A consequence of all this would be that indirect compatibilism is not better, conceptually, than direct compatibilism due to the costs associated with the need to revise our concept. With that said, I think indirect compatibilism would be at least as good as direct compatibilism conceptually. After all, it still covers all the cases that direct compatibilism covers and allows us to maintain that we have free will.

Still, it's once again an open empirical question whether the folk would 'save' free will by revising their concept of free action to include the conditions required for indirect freedom.

Maybe I am wrong and the majority of folk will happily give up free will entirely, rather than revise their concept of free action. Even if that's true I still think my notion of indirect freedom might turn out to be important to the folk for other reasons. For example, many people think being free is a necessary condition for moral responsibility. But if the only actions that are free are the directly free ones (those caused by the agent's FCMs) then the number of cases of genuine moral responsibility could also turn out to be far fewer than we had supposed. Here, perhaps acting indirectly freely is evidence that the agent acted freely for the purposes of moral responsibility, despite not being evidence of an agent's general capacity of free will. Is the folk notion of free action sufficiently nuanced to distinguish between two different senses of free action for two different purposes: directly free actions that count as evidence for our general capacity of free will and moral responsibility, and indirectly free actions that only count as evidence for our responsibility? I don't know. But perhaps it is. Alternatively, maybe it will turn out that what I have called indirect freedom is not a theory of free action at all, but a theory of the conditions under which agents are morally responsible - and it turns out these things come apart. Perhaps that is right, but these are questions I can't answer at this time.

*4.3.1.4. Premise (2), (3), and conclusion.* Now onto Premises (2), (3) and the Conclusion:

2) [Principle of theory selection] If two accounts are equally virtuous in other respects, but one makes judgments better aligned with out pre-theoretic judgments, we should prefer the account that is better aligned with our pre-theoretic judgments.
3) [Theoretical virtues] Indirect compatibilism and straightforward compatibilism are equally virtuous in other respects.
4) [Conclusion] Therefore, indirect compatibilism is an equally good or better account of our concept of free action than is straightforward compatibilism (from 1, 2, and 3).

[Principle of theory selection], I think should be uncontroversial, especially for my project of providing a new account of our concept of free action. With that said, even conceptual engineers might give some weighting to their theory better aligning with our judgments on cases, as the overlap between the engineered content of the concept and content held as part of the folk concept, plausibly makes it more likely the engineered concept will be taken up by the folk.

Similarly, [Theoretical virtues] seems a perfectly plausible claim. You might wonder whether indirect compatibilism and direct compatibilism are relatively alike with respect to other theoretical considerations. For one, direct compatibilism appears to be simpler than indirect compatibilism. However, the upshot of including the conditions required to account for indirectly free actions is not only better alignment with our judgments for cases of free action in the world, but greater explanatory depth to explain why those judgments are free and do so in a

unified fashion. After all, the conditions required for indirectly free actions are not dissimilar from those found in direct compatibilism, for they are still dependent on FCMs. For the purposes of this argument, all I require is for indirect compatibilism to turn out relatively like direct compatibilism with respect to other theoretical virtues and this seems entirely plausible.

#### 4.3.2. Conclusion: Why indirect compatibilism?

Thus, indirect compatibilism is a better (or at worst, as good an) account of our concept of free action than is direct compatibilism. That's because indirect compatibilism provides a better account of our free action judgments than does direct compatibilism. Without the conditions required to successfully account for cases of indirectly free actions, direct compatibilism fails to account for the majority of our free action judgments. Further, even if the conditions required to pick out indirectly free actions are not currently part of our concept of free action, we should expect people to make a principled revision towards including those conditions. After all, the alternative is to discard the majority of cases we might judge to be free in the world and in doing so discard any opportunity that we might possess free will as well.

## 4.4. Conclusion: Indirect freedom

According to indirect compatibilism an agent doesn't just act freely when they perform an action caused by their FCMs, nor, just when they are creating, destroying, triggering, monitoring, and updating action procedures. They also act freely when they perform an action caused by a CAP. The evidence from the brain sciences that I have discussed in this chapter, and previously in Chapter 3, appears to be entirely consistent with this account. More specifically, while there are quite likely some directly free actions there are many more indirectly free actions. In virtue of this, indirect compatibilism is better than direct compatibilism as an account of concept of free action. That's because indirect compatibilism provides a much better account of our free action judgments. Even if the conditions required to pick out indirectly free actions are rejected by the folk, we should nevertheless expect the folk to revise their concept so as to include these conditions. The alternative is to discard the majority of cases we judge to be free and in doing so discard the possibility that we might possess free will. In the next chapter I will discuss some interesting objections to indirect compatibilism. Further, I will provide one piece of evidence from my own cognitive neuroscience investigations that I think shows people updating their action procedures.

## Chapter 5
## Objections and replies to indirect compatibilism

In Chapter 4 I described in detail, and argued for, my new account of free action: indirect compatibilism. In this chapter I will describe and respond to what I think are some of the most interesting objections to my account. Before I begin doing this, though, I think it will be helpful to remind ourselves of some key definitions of terms relevant to indirect compatibilism and briefly recap how they all fit together on my view. As you may recall, according to indirect compatibilism actions are free *simpliciter* if and only if they are caused by a *freedom-conferring mechanism* or a *controlled action procedure*. In particular, they are directly free if they are caused by a freedom-conferring mechanism, and indirectly free if they are caused by a controlled action procedure.

Let's remind ourselves of some of definitions of some of these components.

> An **action procedure** is a sub-personal mechanism that causes action or influences other action procedures in a *contained* and *reliable* fashion.

For example, consider a simple action procedure that causes a patting action. The action procedure will only cause actions of the patting kind (contained) and will rarely fail to cause pats when activated (reliable).

> Something is a **controlled action procedure** iff it's an action procedure that's at least in part under our *direct control*. This means it was either *created*, *destroyed*, *triggered, monitored,* or *updated* by a freedom-conferring mechanism or another controlled action procedure.

Actions caused by controlled action procedures are *indirectly free*.

Let me illustrate with a simple example. Consider a basic controlled action procedure that unconsciously automates and causes wrist-flex actions during a Libet-style experiment. The controlled action procedure only causes actions of the wrist-flex kind and very rarely fails to cause actions of this kind when it is activated. It is thus contained and reliable. This action procedure is a controlled action procedure because it was created by freedom-conferring mechanisms to unconsciously perform the experimental task. For instance, perhaps the person consciously deliberated over the experimental task requirements and as a result of their conscious deliberations caused the action procedure. Creating the action procedure to perform a wrist-flex in response to the experimental stimuli is a directly free action. Further, all the actions caused by this controlled action procedure over the course of the experiment are indirectly free actions.

According to my definition of a controlled action procedure, all that's required for an action procedure to be a *controlled* action procedure is for at least *one* of the direct control conditions - being *created*, *destroyed*, *triggered, monitored,* or *updated* by a freedom-conferring mechanism or another controlled action procedure - to be fulfilled. This means there are a numerous ways action procedures can be controlled action procedures and a wide taxonomy of controlled action procedures to explore. For example, an action procedure is a controlled action procedure if it's created and could be destroyed, triggered, monitored, and updated by freedom-conferring mechanisms or another controlled action procedure. Equally however, an action procedure is a controlled action procedure if it was *not* created, but could be destroyed, triggered and updated by freedom-conferring mechanisms or another controlled action procedure. Unfortunately, I don't have the space in this thesis to explore every species of controlled action procedure. As a result, to simplify matters going forward let me first broadly cleave controlled action procedures into two varieties: *paradigmatic controlled action procedures* and *non-paradigmatic controlled action procedures*.

> Something is a **paradigmatic controlled action procedure** iff it's a controlled action procedure that fulfils all the direct control conditions. This means it is created, and could be destroyed, triggered, monitored, and updated by a freedom-conferring mechanism or another controlled action procedure.

I will say that actions caused by paradigmatic controlled action procedures are *paradigmatically indirectly free*.

> Something is a **non-paradigmatic controlled action procedure** iff it's a controlled action procedure that does not fulfil at least one of the direct control conditions. This means the action procedure was either not created, could not be destroyed, triggered, monitored, or updated by a freedom-conferring mechanism or another controlled action procedure.

I will say that actions caused by non-paradigmatic controlled action procedures are *non-paradigmatically indirectly free*.

To this point in the thesis I have only dealt with paradigmatic controlled action procedures. In this final chapter I shift the focus of my discussion to non-paradigmatic controlled action procedures (henceforth NP-CAP). One reason to discuss NP-CAPs is because I think it likely that they are the most common kind of controlled action procedure. This is because I think that cases where we have the capacity to create, destroy, trigger, monitor, and update action procedures are relatively rare. Another reason to discuss NP-CAPs is because they might be used to motivate *prima facie* arguments against indirect compatibilism. Unfortunately, I

don't have the space in this chapter to consider every species of NP-CAP; instead I will focus on what is potentially the most problematic species of NP-CAP, the **non-deliberative cause**. Roughly characterized for now, non-deliberative causes are controlled action procedures that were created by a freedom-conferring mechanism or some other controlled action procedure, but cannot be destroyed, triggered, monitored, or updated.

The *prima facie* reason non-deliberative causes are problematic is because they cause actions, and can influence other action procedures, in a manner that we can't intercede on at all. For many, the actions caused by non-deliberative causes are likely to be judged unfree. However, according to indirect compatibilism the very fact that we created the non-deliberative cause is enough for the actions they cause to be indirectly free. Though I can't argue for it here, I think that non-deliberative causes are the most potent kind of putative counterexample to indirect compatibilism. Hence if I can motivate the idea that actions caused by non-deliberative causes are indirectly free, then I think I can make the case (though I shall not do so here) that other kinds of NP-CAPs do not pose an issue for indirect compatibilism.

With all this in mind, here is the plan for this chapter. In §5.1 I will provide a complete account of non-deliberative causes. It's only after I have finished giving my account of non-deliberative causes that I can begin working through some of the interesting objections to indirect compatibilism. Each section from §5.2 to §5.5 will present a different objection to indirect compatibilism, along with my reply. In §5.2 I will reply to the objection that non-deliberative causes can't cause free actions because they are psychologically pathological. In §5.3 I will reply to the charge that we aren't morally responsible for actions caused by non-deliberative causes, and thus these actions cannot be free. In §5.4 I will reply to the objection that even if non-deliberative causes cause free actions, they simply can't be free *simpliciter* as they seem completely different in kind to the other free actions caused by a freedom-conferring mechanism. In §5.5 I will reply to the serious charge that I have turned to the dark side and have engaged in nothing but *a priori* theorising. Finally, in §5.6 I will conclude the chapter and in §5.7 I will conclude the thesis.

## 5.1. Non-deliberative causes

Before I begin describing and replying to some interesting objections to indirect compatibilism, I need to provide an account of non-deliberative causes. Remember that the reason I am doing this is because many of the objections that I will be replying to shortly are best motivated by cases of non-deliberative causes, perhaps the most problematic kind of NP-CAP. Let's begin with a definition of a non-deliberative cause:

> Something is a **non-deliberative cause** iff it's a controlled action procedure that only fulfils the direct control condition *created*. This means the action procedure was created, but cannot be destroyed, triggered, monitored, or updated by a freedom-conferring mechanism or another controlled action procedure.

Remember that actions caused by non-deliberative causes are *non-paradigmatically indirectly free*.

Put simply then, a non-deliberative cause is a freely created action procedure that we have *no* ongoing control over. This means that non-deliberative causes cannot be destroyed, triggered, monitored, or updated by a freedom-conferring mechanism or other controlled action procedure. I use the term non-deliberative cause for this, even though there are uncontrolled action procedures over which we have even less deliberative control, because of the special issues they pose for indirect compatibilism.

The easiest way to come to grasp non-deliberative causes is by examples, which I will provide in two parts. First, I will briefly describe a case of a non-deliberative cause that causes actions. After this, I will describe a case of a non-deliberative cause influencing other action procedures. Remember that indirect compatibilism judges that actions caused by non-deliberative causes are free and that the influence of non-deliberative causes on our other action procedures is not freedom-undermining. That's because the non-deliberative cause was freely created by a freedom-conferring mechanism or another controlled action procedure. For ease of explication I will assume indirect compatibilism gets these free action judgments correct for now. I will describe and reply to some important objections from the next section onwards.

### 5.1.1 Non-deliberative causes causing actions

One thing non-deliberative causes can do is cause actions over which we have no control. Let me illustrate with a simple case. Imagine an agent, Bud, who creates a non-deliberative cause that causes him to physically retaliate whenever he is attacked. One evening Bud is socializing with his mates at the pub after a particularly long day at work. Bud doesn't drink, but his friends do, and by the end of the evening they are heavily intoxicated. After making sure that his friends have a safe ride home available, Bud informs his friends that he is going as he is feeling tired and will walk home as it's a nice evening outside. On his way home, along a particularly dark stretch of road, Bud is unexpectedly attacked from behind. Without thought Bud turns and retaliates against his attacker landing a heavy punch square on their jaw. The punch leaves the attacker unconscious on their feet and they hit the ground hard. In that moment, Bud notices that the attacker was one of his heavily intoxicated 'friends' whom he had just been drinking with earlier. Bud regrets what he has done and wishes that he had never created the non-deliberative cause that makes him physically retaliate whenever he is attacked.

Perhaps Bud wishes that he could have deescalated the situation by talking down his attacker, subduing them, or fleeing the scene all together. However, none of these actions were open to Bud when he was attacked. Once a non-deliberative cause has been created, the agent no longer has control over said cause, causing actions. Bud freely created a non-deliberative cause that makes him physically retaliate when he is attacked and this means that when he is attacked the non-deliberative cause causes him to retaliate no matter what. According to indirect compatibilism, Bud's retaliatory strike in this case is an indirectly free action and that is because Bud freely created the non-deliberative cause that causes him to physically retaliate whenever he is attacked. If you think indirect compatibilism gets these judgements wrong, that's fine; in the coming sections I will be discussing and replying to what I think are some of the best objections to my view. For now, I just want to describe some non-deliberative causes and explain why indirect compatibilism judges that they cause free actions.

### 5.1.2 Non-deliberative causes influencing other action procedures

Of course non-deliberative causes don't just cause actions; they can also influence other action procedures. Once again, let me illustrate with a simple case. Imagine an agent Bodi who creates a non-deliberative cause to prevent him from killing other people. Bodi is a police officer tasked with patrolling some rural stretches of road that surround the city. One evening while he is patrolling he notices that the traffic barrier on one of the bends has been broken. He pulls his patrol car over, holsters his gun and begins to investigate. Beyond the broken traffic barrier and down a bank he finds a rolled truck with a local driver still trapped inside. The driver is conscious and does not appear to be badly injured. While Bodi tries to free the trapped driver he quickly realizes that it will be impossible without assistance. Bodi begins casually conversing with the truck driver to keep him calm while they wait for assistance, but as they converse, a small fire catches at the back of the truck and begins spreading rapidly. The truck driver sees the approaching fire and, knowing he's about to die, turns to Bodi and begs him to not let him suffer needlessly by burning to death. Bodi breaks down, for he too knows the truck driver is going to die and doesn't want him to suffer by burning to death, but there is nothing he can do to prevent the truck driver from dying in this manner. Still, he realizes that if he could kill the truck driver then the driver would not have to suffer needlessly. In that moment Bodi regrets creating the non-deliberative cause that prevents him from killing other people.[43]

[43]This case is adapted from a case in Hart's (1968) *Punishment and Responsibility*. Some readers will be familiar with this kind of case from ongoing debates around the moral status of killing and letting die (e.g., Rachels, 1975; Kamm, 2007; Woollard, 2015). While I think there is no significant moral difference between the acts of killing and letting die, nothing hangs on this fact in the thesis.

Once the non-deliberative mental cause has been created, the agent no longer has control over its influence on other action procedures. To be clear, I don't think the only way non-deliberative causes could influence other action procedures is by removing them as options from the agent. In some cases non-deliberative causes could make it more, or less, difficult to choose some action procedures as options. Still, I will focus on cases where non-deliberative causes *remove* action procedures as options as these cases are potentially more problematic for indirect compatibilism.

According to indirect compatibilism Bodi acts freely when he chooses to comfort the truck driver as best he can, even though he really wanted to kill the truck driver to prevent him needlessly suffering.[44] That's because it's impossible to do something that is not an option for you do.[45] I might really want to fly, but the fact that I can't fly is not freedom undermining as it was not something I could do prior to being in the choice situation. Similarly, Bodi's acting freely cannot be undermined by not being able to kill the truck driver, as this is something he could not do prior to being in the choice situation. Non-deliberative causes can allow agents to remove certain action procedures from their action sets, making them psychologically impossible to do. Once again you might think indirect compatibilism gets this judgment wrong and that's fine, I will begin discussing objections and replies to my view shortly. All I'm doing at the moment is describing non-deliberative causes influencing other action procedures and explaining why nothing freedom undermining occurs, *according to indirect compatibilism*, when they influence other action procedures.

Before moving on, I will briefly outline some of the different ways non-deliberative causes could influence other action procedures. I think that my account becomes much more concrete if I can outline some of the possible ways that non-deliberative causes might influence our other action procedures. Whether or not any of the suggestions are in fact realized in beings like us is not something I will address in this thesis, but is a task for my future self and the various brain sciences. The options I will briefly outline are: *information screening*, *deliberation screening*, *choice screening* and *action screening*.

5.1.2.1. *Information screening*. One way non-deliberative causes can influence other action procedures is by information screening. This means that the non-deliberative cause selectively filters the information we have available, and in doing so makes some action procedures more, or less, attractive as options than other action procedures. In some cases the non-deliberative cause might selectively filter all the information about a particular action procedure causing it to

[44]Remember that on my view, free will is *not* about "freedom to do *X*". Free will is about when we do act, do we do so in a manner that is *reliably* free.

[45]See David Braddon-Mitchell (forthcoming) *Freedom and binding consequentialism*.

not appear as an option to the agent. For example, one reason Bodi might think there are no other options available to him aside from comforting the truck driver is because he was *sincerely* unaware of the gun holstered on his side. When the non-deliberative cause filtered out this crucial piece of information, it meant Bodi had no good reason to think that the option to shoot the truck driver was available to him. And if Bodi had no good reason to think that the option to shoot the truck driver was open, then he had no good reason to do it.

5.1.2.2. *Deliberation screening.*[46] Another way non-deliberative causes can influence our action procedures is by influencing which action procedures we can deliberate over. Crucial, here, is that not all information, or options we are consciously aware of, are ones which actually enter the causal process of deliberation and action selection. The action procedures that we deliberate over represent the pool of action procedures that we can freely choose from when deciding what to do. If an action procedure is not part of the deliberation pool then it's not an option for choice, even if we can entertain the idea, and there are no constraints on what we can choose within the choice set. For example, Bodi might be aware of the fact he has a gun holstered on his side and he might judge that if he kills the truck driver he will prevent the driver from suffering needlessly. But if the non-deliberative cause has filtered out the killing action procedure from his deliberation pool, then the option to kill the truck driver was never there for Bodi to choose, even though he can entertain the idea of killing the truck driver.

5.1.2.3. *Choice screening.* Closely related to deliberation screening, another way non-deliberative causes might influence our action procedures is by influencing which action procedures we can actually choose. For example, Bodi might be aware of the option of killing the truck driver and the option to choose the killing action procedure might be in his deliberation pool, but the non-deliberative cause causes him to be unable to choose it. Instead, the best option that he can choose is to comfort the truck driver. The key difference between choice screening and deliberation screening is that while in choice screening there is an apparent option open to choose, in deliberation screening there is not even the apparent option. The intervention is at a later stage: at the point of choice, rather than at the point of choosing options to (seriously) deliberate over.

5.1.2.4. *Action screening.* The final way a non-deliberative cause can influence our action procedures that I will outline is by influencing which actions action procedures can actually cause. For example, Bodi might be aware of the option of killing the truck driver, find the option to choose the killing action procedure is in his deliberation pool and be able to choose the killing action procedure. But after he has chosen the killing action procedure, Bodi finds that the killing

[46]See David Braddon-Mitchell (forthcoming) *Freedom and binding consequentialism.*

action procedure cannot cause killing actions. He might still be able draw his gun, aim at the truck driver's head, rest his finger firmly on the trigger, and so on. However, the non-deliberative cause in this case actively filters the actions that are being caused by the killing action procedure so that no matter what Bodi cannot actually kill the truck driver.[47]

### 5.1.3 Conclusion: Non-deliberative causes

In this section, I have provided and illustrated some examples of non-deliberative causes. For now I have simply taken for granted that indirect compatibilism is correct and that the actions caused by these causes are free; that the actions that occur as a result of the influence of non-deliberative causes are no less free than those that occur without their influence. In the next section, I will begin the process of explaining and replying to some objections to indirect compatibilism that arise from consideration of non-deliberative causes. Perhaps the most obvious is the one which I will discuss first; the objection that actions from non-deliberative causes cannot be free as they reflect a psychological pathology. Since indirect compatibilism says these actions are free, this presents an objection to indirect compatibilism.

## 5.2. Objection 1: Non-deliberative causes are pathological

Now that we have a grasp of non-deliberative causes let's begin working through some of the most interesting arguments against indirect compatibilism that consideration of non-deliberative causes can be used to motivate. Indirect compatibilism judges that actions caused by non-deliberative causes are free, and their influence on other action procedures is not freedom undermining. But if we can identify some non-deliberative causes that cause unfree actions, or influence other action procedures in a freedom undermining manner, then we have a counterexample to indirect compatibilism. This will be the basic strategy of the first objector.

The first objector's charge is that non-deliberative causes cannot issue in free actions because non-deliberative causes are pathological. That's because anything which causes you to perform non-maximising actions without the possibility of intervention is pathological.[48] Actions

[47]It's plausible that each of these mechanisms cause you to have different intuitions about free action. I take it that some variance in judgments for cases of free action in the literature cases could be explained by people reading one of these mechanisms into the case. For example, information and deliberation screening are *prima facie* compatible with our free action because the non-deliberative cause doesn't influence our capacity to choose directly. Action screening is less about our free action and more just a defeater to our freedom *simpliciter*. The capacity to choose is not directly affected, rather just like cases of physical restraint you are not free to act in the way that you choose. Choice screening might give us pause for concern as our capacity to choose is being influenced directly. I think it's possible that thoughts about something like choice screening might have motivated the original challenge of the brain sciences. The earliest signs of predictive neural activity were often referred to as the 'choice', even though participants were unconscious of their 'choice' at this time. That's because the neural activity was significantly associated with what participants would go on to report choosing. Credit and thanks to David Braddon-Mitchell for this helpful suggestion.

[48]It is contentious, even amongst contemporary naturalistic accounts of disease, what makes something pathological. However, one thing in common between these different accounts is that pathology results in the impairment of

caused by pathologies are unfree. Hence actions caused by non-deliberative causes are unfree. Here's the basic motivation behind the objection. Consider a paradigm mental pathology, obsessive compulsive disorder (OCD), which is characterized by the sufferer's inability to control certain repeated actions, or to influence repeated intrusive thoughts.[49] Part of what makes OCD pathological is the fact that the sufferer acts in a non-maximising manner without the possibility of being able to intervene on their actions or their intrusive thoughts at the time of action. If non-deliberative causes are what realize mental pathologies like OCD, then indirect compatibilism judges that people in the grip of pathologies like OCD act freely. But if we believe that the actions caused by mental pathologies are unfree, then indirect compatibilism must be false.

Here, then, is the first objector's objection:

Objection 1

1) [Pathology non-maximising] Anything which causes you to *reliably* perform acts which are (a) non-maximising, by your lights, at the time of the action and (b) for which there is no possibility of intervention, is pathological.
2) Some non-deliberative causes cause you to *reliably* perform acts which are (a) non-maximising, by your lights, at the time of the action and (b) for which there is no possibility of intervention.
3) [Lemma] Therefore, some non-deliberative causes are pathological (from 1 and 2).
4) [Pathology unfree] All actions caused by pathologies are unfree.
5) [Conclusion] Therefore, some non-deliberative causes cause unfree actions (from 3 and 4).

The argument is valid. In this section, I will attempt to undermine [Pathology non-maximising] and [Pathology unfree]. I will begin by undermining [Pathology unfree]. I will argue that either the pathology that is causing actions is *not* a non-deliberative cause, but just an action procedure, in which case indirect compatibilism has no problem judging the actions it causes are unfree, or the pathology that is causing the actions is a non-deliberative cause in which case the actions it causes are free, and ought to be judged to be free. Remember that built into the direct control condition is that to be a non-deliberative cause requires that the action procedure (which is the non-deliberative cause) does what it was created to do. If the pathology does not fulfil the direct

---

normal functional ability (e.g., Boorse, 1977; Wakefield, 1992; Griffiths, 1993; Godfrey-Smith, 1994). The fact that non-deliberative causes cause people to perform below typical efficiency, by making them perform non-maximizing actions without any possibility of intervention, suggests, *prima facie*, that they are pathological.

[49]See Abramowitz, Taylor and McKay (2009) for an excellent overview on obsessive compulsive disorder. Obsessive compulsive disorder is not the only example of a mental pathology that could 'appear' like a non-deliberative cause. Numerous mental pathologies, such as eating, anxiety and panic disorders are centred on the inability of the sufferer to control certain actions, or the influence of certain intrusive thoughts.

control condition, then it's *not* even a non-deliberative cause and the actions it causes are unfree, even according to indirect compatibilism. If, on the other hand, the pathology *is* a non-deliberative cause, then the actions it causes are free. That's because the non-deliberative cause is doing exactly what it was freely created to do. But, I will argue, in these cases it is not so counterintuitive to suppose that the actions in question are indeed free. All in all, I will argue that some, but not all actions caused by pathologies are unfree, by making the case that those pathologies that are non-deliberative causes do indeed cause free actions.

After this, I will undermine [Pathologies non-maximising]. Here, then, is my argument that some non-deliberative causes are not pathological because possessing them is globally maximising:

Non-pathological non-deliberative causes

1) [Not pathology] Anything which is globally maximising to possess is not pathological.
2) [Global maximising] Some non-deliberative causes cause you to *reliably* perform acts which are (a) non-maximising, by your lights, at the time of the action, (b) for which there is no possibility of intervention and (c) are globally maximising to possess.
3) Therefore, those non-deliberative causes that are globally maximising to possess are not pathological (from 1 and 2).

The argument is valid. Is it sound? I think so. Even if a non-deliberative cause causes you to act in a manner that is non-maximising at the time of action, the overall benefits of being able to bind yourself to perform certain actions, and refrain from performing others, might be far greater. Something doesn't count as pathological if, in general, under usual circumstances, having it is more beneficial than not having it.[50] If the possession of a non-deliberative cause is globally maximising, then even if it sometimes causes acts that are locally non-maximising, the non-deliberative cause is not pathological. If the non-deliberative cause is not pathological, then we have no reason to suppose that the actions caused by the non-deliberative cause are not free. I'll demonstrate this claim by outlining how some non-deliberative causes can cause people to act in a more pro-social and cooperative manner towards others. If acting this way is globally maximising, then this is an example of non-deliberative causes that are globally maximising, whilst still sometimes being locally non-maximising.

### 5.2.1 Non-deliberative causes and *mere* action procedures

Let's start with a case of mental pathology that *prima facie* looks like a non-deliberative cause that causes unfree actions and influences other action procedures in a freedom undermining manner.

[50]In many cases it could be vague whether or not possessing some non-deliberative cause is globally maximizing. I will have much more to say on vagueness and its relationship to our judgments for free action in §7.4.

Imagine that Andrew creates an action procedure that causes him to delete typographical errors in his thesis as he proof reads. Things are fine at first; Andrew scans each page and when he comes across a mistake he performs a deleting action and makes his corrections. Sometime later things start to go wrong; Andrew finds himself deleting text in the thesis even when he knows there are no mistakes. He also finds himself having recurring intrusive thoughts that there are mistakes on every page of the thesis that need deleting. Andrew concludes that his deleting action procedure is not doing what it was created to do and attempts to update and fix it. When that fails he tries destroying it and that fails too. Much to his dismay, Andrew realizes that he no longer has any control over the haywire deleting action procedure.[51]

This case *roughly* approximates a case of OCD. I will assume for the moment that [Pathology unfree] is true, and Andrew's actions are unfree because he is in the grip of a mental pathology. According to this first objector, the action procedure at the centre of Andrew's OCD is a non-deliberative cause. Andrew *created* the deleting action procedure that he subsequently has *no* control over. He cannot control his deleting actions, or control the influence of his deleting thoughts. He is unable to update or destroy the deleting action procedure. At the very least, it seems to be something like a non-deliberative cause. Yet we are inclined to say that Andrew's actions are not free. Hence we have a counterexample to indirect compatibilism: an action that the view categorises as free, but which is not free.

I concede that this action is not free. However, it's not a counterexample to indirect compatibilism because the action procedure at the centre of this case is not a non-deliberative cause at all. Andrew created an action procedure to delete any typographical errors he came across while proof reading his thesis. He did not create an action procedure that would cause him to delete the good text in his thesis or cause recurring intrusive thoughts about deleting. The fact that Andrew's deleting action procedure is not doing what it was created to do signals that the action procedure is not a non-deliberative cause. That's because it is built into the creation condition that to be a controlled action procedure, that action procedure does what it was created to do. So while the mechanism in question does count as an action procedure, it doesn't count as a controlled action procedure. But indirect compatibilism is not the view that actions caused by action procedures are free: it's the view that actions caused by **controlled** action procedures are free, and this is not a case in which the action in question is caused by a

[51]This is only meant as an illustrative case and nothing that I say hangs on the specifics of the case. For example, you might think that Andrew does have the ability to destroy, trigger off, or update this dysfunctional action procedure. After all, people do overcome their OCD, or at the very least learn to manage it. I think this is right, but it still seems that people very rarely overcome something like OCD without professional assistance and psychoactive medication. Thus, while it is not impossible to think that Andrew could destroy, trigger or update the dysfunctional action procedure it is very unlikely.

controlled action procedure. Hence, indirect compatibilism, correctly classified this as a case of unfree action.

To illustrate this point, consider another person, Andrew*, who freely creates a non-deliberative cause that is exactly like original Andrew's deleting action procedure. That is, Andrew* freely furnishes his psychology with the OCD pathology. Once again, I am assuming for the moment that [Pathology non-maximising] is true and that anything that causes you to perform non-maximising actions without the possibility of intervention is pathological. Conditional on there be nothing wrong with Andrew* that causes him do this, according to indirect compatibilism when Andrew* goes on to delete the majority of his thesis he acts freely. Why? Because, the action procedure is in fact a non-deliberative cause: it does exactly what it was created to do. In this case it seems reasonable to think that Andrew* has created a non-deliberative cause that is a mental pathology, and yet the actions that issue from that non-deliberative cause are free actions.

To recap, indirect compatibilism has no issue judging that original Andrew is unfree because the action procedure in this case is not a non-deliberative cause because it is not doing what it was created to do. On the other hand, indirect compatibilism correctly, to my mind, judges that Andrew*'s action is free because in effect he freely created a mental pathology.[52]

Indirect compatibilism, then, only judges those actions which arise in virtue of mental pathologies to be unfree when they are caused by action procedures that fulfil none of the direct control conditions to be a CAP. But if someone freely creates a mental pathology then it seems *plausible* that the actions that result from that pathological non-deliberative cause are free. If I am right then [Pathology unfree] is false, and some actions issuing from pathological non-deliberative causes are free. Next, I will undermine [Pathology non-maximising].

### 5.2.2 Non-deliberative causes, human pro-sociality and cooperation

Here is the argument I outlined earlier, which if sound, shows that some locally non-maximising non-deliberative causes are not pathological.

Non-pathological non-deliberative causes

1) [Not pathology] Anything which is globally maximising to possess is not pathological.

[52]Once again nothing I say here hangs of the specifics of this case. One intuition I have about case Andrew* that I want to flag is that I think the actions caused by his non-deliberative cause are *less* free relative to actions caused by freedom-conferring mechanisms and other controlled action procedures. For now, I am sticking with the current orthodoxy of the free will literature that a *necessary* condition for being free is that it is all or nothing. Shortly, in §7.4 I will outline how my account of indirect compatibilism naturally admits for the possibility of their being degrees of freedom and as a result vindicate my intuition about the status of Andrew*'s freedom.

2) [Global maximising] Some non-deliberative causes cause you to *reliably* perform acts which are (a) non-maximising, by your lights, at the time of the action, (b) for which there is no possibility of intervention and (c) are globally maximising to possess.
3) Therefore, those non-deliberative causes that are globally maximising to possess are not pathological (from 1 and 2).

The basic idea is that even if a non-deliberative cause causes you to sometimes perform actions that are *locally* non-maximising, possession of the non-deliberative cause could still be *globally* maximising. Provided we also think that anything which is globally maximising does not count as pathological, then [Pathology non-maximising] is false, and some locally non-maximising non-deliberative causes are not pathological.

If it's globally maximising to possess some non-deliberative causes, then they are exactly the kinds of things that we should predict to find in beings like us. Here I will outline just one line of evidence that shows [Globally maximising] is true. Binding yourself psychologically to perform certain actions, or to refrain from performing others, can be incredibly beneficial, even if it can often result in performing locally non-maximising actions. Specifically, I will use the case of strong reciprocity, where an individual psychologically binds herself to punish and cooperate even when it is costly to them to do so, because binding themselves in this manner makes them a more attractive person to cooperate with. While I think these benefits are shared with other group members, here I will focus solely on the benefits to the individual who possesses the non-deliberative cause.

Reciprocity is taken to be the general mechanism that accounts for evolution and maintenance of pro-sociality and cooperation amongst large groups of unrelated individuals.[53] Roughly, reciprocity describes the tendency to repay good acts with good acts and bad acts with bad acts when dealing with other people. Here I will briefly outline two kinds of reciprocity in the literature, *weak* and *strong*, and will suggest that *strong* reciprocity could be caused by non-deliberative causes. If it's globally maximising to become a strong reciprocator in some cases, then we should expect to find people who possess non-deliberative causes. Further, provided you also think that anything which is beneficial, under usual circumstances, is not pathological then [Pathology non-maximising] is false, and some locally non-maximising non-deliberative causes are not pathological.

Let's start with *weak* reciprocity.[54] Someone employing weak reciprocity will punish and cooperate when it is beneficial for them to do so. Strictly speaking someone employing weak

[53]Cooperation between related individuals is explained by the theory of inclusive fitness (Hamilton, 1964; for review see Gardner & West, 2014).
[54]*Weak* reciprocity is also commonly referred to in the literature as reciprocal altruism.

reciprocity could elect to forego punishing when it is too costly, or not cooperate when it would be more beneficial for them to defect. I will not discuss weak reciprocity in any more detail as it is not a strategy that I think could be realized by a non-deliberative cause (see seminal papers by Axelrod, 1984; Gouldner, 1960; Trivers, 1971; for recent review see Nowak, 2006). It would be better realized by a controlled action procedure that can be triggered on and off as required. In contrast, someone employing strong reciprocity will punish and cooperate even when it is not locally beneficial for them to do so (for example see, Boyd, Gintis and Bowles, 2010; Fehr and Fischbacher, 2003; Henrich, Boyd, Bowles, Gintis, Camerer, & Fehr, 2004). Hence it makes sense that such behaviour would be the result of a non-deliberative cause.

Following Francesco Guala (2012) I will call the action of costly punishment *negative* strong reciprocity and the act of costly cooperation *positive* strong reciprocity. My suggestion is that one way we can make sense of why people perform actions that appear to be locally non-maximising at the time of action, is because they are being caused by a globally maximising non-deliberative cause.[55] I will address both components of strong reciprocity next, as they correspond nicely to two ways a non-deliberative cause might function. Roughly, for now, the reason it is globally maximising, under usual circumstances, to be a strong reciprocator is because it causes others to interact with you in a much more prosocial and cooperative manner.

5.2.2.1 *Negative strong reciprocity.* The easiest way to grasp how being a negative strong reciprocator can cause others to interact with you in a more prosocial and cooperative manner is by way of example. Remember, I am doing this in in order to illustrate how it can be globally maximising to have a non-deliberative cause that binds and causes actions that you cannot intercede on.

Imagine you and a stranger are invited to play a one-off game. The rules of the game are simple; you will receive $100 and be asked to divide the money between you and the stranger. You can offer the stranger any amount, and if they accept you will each receive the amount contained in the offer. However, if the stranger rejects the offer you will each receive nothing. After hearing the rules, you decide to offer $1 to the stranger and keep $99 for yourself. The stranger automatically rejects the offer and you both receive nothing.[56]

---

[55]I am not claiming that *all* cases of strong reciprocity are caused by freely created non-deliberative causes. Some cases might be as some evolutionary psychologists tell us, part of our psychological tool-box evolutionarily selected for in the environment for evolutionary adaptation (e.g., Tooby & Cosmides, 1992). Here is why I think *some* cases are caused by freely created non-deliberative causes. First, there is evidence that the propensity to engage in strong reciprocity is part be contingent on their being strong group norms. For example, the Tsimane of the Amazon make more unfair offers and do not routinely reject unfair offers because they appear to have no strong group norms regarding the distribution of resources (e.g., Gurven, Zanolini, & Schniter, 2008). Second, there is evidence that engaging in strong reciprocity in one area does *not* predict engaging in strong reciprocity in another area (e.g., Yamagishi et al., 2012). Together these pieces of evidence together suggest that at least some cases of strong reciprocity are caused by freely created non-deliberative causes.

[56]There is plenty of evidence that people engage in costly punishment of this sort both in the behavioural laboratory (see for example, Güth, Schmittberger, Schwarze, 1982, Fehr, & Fischbacher, 2005; Gächter, & Herrmann, 2008)

For many people the stranger's rejection of the offer is entirely understandable, but it also makes little sense. To see why it makes little sense, we need do a bit of basic game theory. You make the first move you and must decide whether or not to make a fair offer (50/50) or an unfair offer (99/1). Assume for the moment that the stranger will automatically accept any fair offer, then if you make a fair offer the game ends with you each $50 up. Still, if you make an unfair offer you stand to gain more than if you make a fair offer. If you make an unfair offer, then the stranger has a choice to make, either they choose to reject the offer and receive nothing or choose to accept the offer and receive $1. The stranger's choice is between receiving $0 and receiving $1. Rationally speaking the stranger should choose the $1. The solution to this game is that you make an unfair offer and the stranger accepts. But that can't be right! After all, most of us think that the stranger's rejection of the unfair offer made complete sense. We are missing something important in this simple case.[57]

What if the stranger in this case were able to signal that they would as a matter of fact reject any offer that is not a fair offer? Suddenly the original player's choice is no longer between receiving $50 or $99. It's now between receiving $50 or $0, and no one would choose $0. By signalling that they would reject any unfair offer, the stranger is also signalling that they will make a locally non-maximising action. One effective strategy to game someone into acting in a pro-social and cooperative manner is to convince them that you will perform a locally non-maximising action and will punish them even when it is not beneficial for you to do so.[58]

Let's tie this back to my argument about why some locally non-maximising non-deliberative causes are not pathological. One way that someone can psychologically bind themselves to punish all unfair offers is with a non-deliberative cause. So long as the stranger is guaranteed to punish all unfair offers, this causes the player to have to choose between being cooperative and receiving something or be uncooperative and receiving nothing. Under usual circumstances, a non-deliberative cause that causes you to punish all unfair offers is globally maximising because it causes others to have to act cooperatively if they want to choose the maximising action. Even if the non-deliberative cause causes you to act in a locally non-maximising manner whenever people defect (don't cooperate), overall the non-deliberative cause will cause others to act towards you in a much more prosocial and cooperative manner. If this is right then [Globally maximising] is true. Then, provided you think that things which are globally

---

and cross-culturally (see for example, Roth, Vesna, Masahiro, & Shmuel, 1991; Cameron, 1999; Mathew, & Boyd, 2013). Though, see Guala (2012) for a critical discussion on the scope of these kinds of evidence.

[57]In a one-off game the weak reciprocator should accept the unfair offer. There is nothing beneficial about adopting a punishing strategy with someone you will not interact with again.

[58]This line of thinking can be used to explain phenomena from blood feuds and vendettas (e.g., Ericksen, & Horton, 1992; Nikiforakis, & Engelmann, 2011; Fehl, Sommerfeld, Semmann, Krambeck, & Milinski, 2012), and nuclear deterrence theory (e.g., Rauchhaus, 2009; Quackenbush, 2010; Sechser & Fuhrmann, 2013).

maximising don't count as pathological, the non-deliberative cause in this case is not pathological.

Let me briefly address a couple of concerns. First, I have stated that non-deliberative causes can be globally maximising under usual circumstances. It's possible to imagine creating a non-deliberative cause, which binds you to perform certain actions, or away from being able to perform certain actions which is globally maximising in one environment but not globally maximising in another environment. Perhaps there is some sudden dramatic shift in the way that group members interact with one another that causes the non-deliberative cause to be maladaptive. My thoughts on this case are the same as in §5.2.2. The fact that the non-deliberative cause is no longer globally maximising could cause it count as being pathological. However, just because some non-deliberative cause counts as being pathological doesn't mean the actions it causes are unfree. Maladaptive non-deliberative causes are not dysfunctional as they are doing exactly what they created to do.[59] Even if it turns out that in some cases [Pathology non-maximising] is true, provided [Pathology unfree] is false then the actions caused by the non-deliberative cause are still free.

Second, you might think that it's possible to simply *fake* the fact that you are a strong reciprocator. If this were possible you could receive the overall benefits of being a strong reciprocator without psychologically binding yourself to have to perform locally non-maximising actions. Using the continued example of unfair monetary offers, to maintain the ruse you would have to continuously choose to do the locally non-maximising thing every time someone defects on you. There is, however, plausibly a discernible phenotypic difference between someone merely signalling that they will respond in a locally non-maximising manner to various group norm violations and someone that is signalling that they are psychologically bound to respond in a locally non-maximising manner to various group norm violations.[60] The best way to signal that you are strong reciprocator to unfairness, killing, and so on is to actually make your psychology such that you will perform costly punishment on anyone who performs those actions. After all, the reason that strong reciprocators are so effective at causing others to act in a pro-social and cooperative manner towards them is because their commitment to various group norms is unwavering.[61]

---

[59]The idea that something is functional and counts as being pathological due to a mismatch with the environment comes from Griffiths and Matthewson (2018) and Matthewson and Griffiths (2017).

[60]Creating a non-deliberative cause to enforce pro-sociality and cooperation might be a form of *costly* signalling. This means that is must be hard to fake to other group members. See for example, Smith, & Bird (2000); Gintis, Smith, & Bowles (2001); Lotem, Fishman, & Stone (2003).

[61]One final piece of evidence that might sweeten my suggestion here comes from neuroimaging work on costly punishment (which is called effective punishment in this literature). One key brain location active during costly punishment is the dorsal striatum, a brain region associated with reward processing during goal-directed activities.

5.2.2.2 *Positive strong reciprocity.* Non-deliberative causes don't just cause actions, they can also influence other action procedures, and in some cases this influence can remove options that we can choose from. Just as being psychologically bound to perform certain actions can cause people to act in a prosocial and cooperative manner, so too can being psychologically bound to be unable to perform certain actions. If you are able to signal to people that you are incapable of performing actions that violate various group norms then you make yourself a more attractive group member and cooperative partner. For example, I think that possessing a non-deliberative cause that removes the option of killing is a common non-deliberative cause that might be found in people's psychology. And people who don't have the option of killing you are more attractive as group members and cooperative partners. While this might result in them being unable to choose the maximising action in some cases, in usual circumstances the overall benefits will be far greater. Once again if this is right then [Globally maximising] is true and provided you think that things which are globally maximising don't count as pathological then the non-deliberative cause in this case is not pathological.

One other plausible upshot of removing certain actions and action procedures as options with a non-deliberative cause is that you remove yourself as a target of strong reciprocity. There are going to be cases where, by some accounts, the right thing to do, is to do the thing that violates some group norm. For example, in the case of Bodi that I described earlier, arguably the right thing for Bodi to do, if the option were available to him, would be to shoot and kill the truck driver before he burned to death. But killing the truck driver would violate norms against killing and cause strong reciprocators to punish Bodi. Very few people think that a person should be punished if they are unable to kill someone even if, on some accounts, killing would be the right thing to do. However, it's plausible that many people think that someone should be punished for killing even if, by some accounts, that person does the right thing by killing.

Do people really punish people for violating a group norm even when they judge that the act of violating a group norm was the right one? This is an open empirical question but there is some interesting suggestive evidence from a set of pilot data reported by Peter Railton (2014; unpublished). Railton presented students with a simple scenario where they were asked whether or not they would shove a corpulent gentleman on top of a possible suicide bomber in order to prevent him from getting on board the bus. If we accept that the number of people who say they would perform this action is indicative of what they think the right action is, then approximately two-thirds of people think the right thing to do is push the corpulent gentleman on top of a

---

Not only is the dorsal striatum active during costly punishment, the strength of activation is proportionate to the cost of the punishment. People who engage in costly punishment are not only doing so with purpose they feel good about what they are doing (de Quervain, et al., 2004).

suicide bomber. However, when asked whether or not they would trust the person who pushes the corpulent gentleman on top of the suicide bomber, the majority of people respond that they would not. Mistrust is, by my lights, a form of punishment, as it will cloud any future interactions that others have with that person. Importantly, for me, insofar as this mistrust constitutes *a* punishment, then we have suggestive evidence of people punishing people for doing what even they think is the right thing. Of course, this is only a pilot dataset and a full confirmatory study is required, but it fits with the account that I have described. Still, I think we should expect people to possess non-deliberative causes not only to cause costly acts of punishment, but to prevent certain action and action procedure options from being able to be chosen.

In summary, I have illustrated that under usual circumstances, it can sometimes be globally maximising to psychologically bind yourself to cause certain actions, or to refrain from producing others. Together with the idea that things which are globally maximising don't count as pathological, this gives us good reasons to think [Pathology non-maximising] is false. As a result, we no longer have any reason to think that non-deliberative causes must be pathological and hence cause unfree actions.

**5.2.3 Conclusion: Non-deliberative causes are pathological**

In this section I have addressed the charge that non-deliberative causes cause unfree actions because they are pathological. The reason non-deliberative causes are pathological is because they often cause actions that appear, at the time of action, to be non-maximising. In this section I have attempted to undermine this argument in two ways. First, even if we accept that something that causes non-maximising actions counts as pathological, there is no reason to think that all pathologies cause unfree actions. That's because there is a difference between pathologies realized by action procedures that fulfil none of the direct control conditions to be a CAP, and non-deliberative causes. Indirect compatibilism has no issues counting actions caused by action procedures that fulfil none of the conditions to be a CAP as unfree. On the other hand, indirect compatibilism counts actions caused by non-deliberative causes as free, even if such a non-deliberative cause still counts as being pathological. I think this is the right judgement because the non-deliberative cause is doing exactly what it was freely created to do.

Second, just because non-deliberative causes sometimes cause locally non-maximising actions without the possibility of intervention doesn't mean they count as pathological. It can be globally maximising to psychologically bind yourself to perform certain actions, or away from being able to perform certain actions. To illustrate this claim I outlined the overall benefits of negative and positive strong reciprocity to the individual. Even though psychologically binding yourself to punish and cooperate means you might often perform locally non-maximising

actions, binding yourself in this manner also causes other people to act in a far more prosocial and cooperative manner towards you. This means that under usual circumstances it is globally maximising to possess the relevant non-deliberative causes. Provided you think that something which is globally maximising is not pathological then we have good reasons to think that some non-deliberative causes are not pathological and hence cause free actions.

### 5.3. Objection 2: Non-deliberative causes and moral responsibility[62]

According to indirect compatibilism, actions caused by non-deliberative causes are free. However, once a non-deliberative cause has been created we have no ongoing control over the actions it causes, or the influence it has on our other action procedures. According to the second objection, because we are unable to intercede on actions caused by non-deliberative causes (or the influence they have on our other action procedures) we cannot be morally responsible for these actions. Moral praise and blame are only appropriate reactions when we *deserve* such reactions and actions not under our control are not deserving of our moral judgements. Importantly, because of the tight conceptual connection between moral responsibility and acting freely, if we are not morally responsible for actions caused by our non-deliberative causes then they are not free. Here, then, is the argument that underlies the second objection.

Objection 2

Consider some standard action *A* which is caused by a non-deliberative cause:

1) If *A* is free then someone, *S*, performing *A* is morally responsible for *A*.[63]
2) *A* is caused by a non-deliberative cause.
3) [Principle of intercession] *S* is morally responsible for *A* only if *S* is able to intercede on *A*.[64]

---

[62]Thanks and credit to Michael Duncan, James Norton and Nathaniel Gan for their assistance with this objection.

[63]There is another way to run this argument. Let's call it the freedom-first route:

1) If *S* performing *A* is morally responsible for *A*, then *A* is free.
2) *A* is caused by a non-deliberative cause.
3) [Freedom to do otherwise] *S* performs *A* freely only if *S* is able to intercede on *A*.
4) [No intercession] If *A* is caused by a non-deliberative cause, *S* cannot intercede on *A*.
5) [Lemma A] *S* cannot intercede on *A* (from 2 and 4).
6) [Lemma B] *A* is unfree (from 3 and 5).
7) [Conclusion] Therefore, *S* is not morally responsible for *A* (from 1 and 6).

Whether you prefer the moral responsibility first route or freedom first route will depend on whether you think moral responsibility or freedom is more foundational. I think that freedom is more foundational, but there are many theorists who don't, such as those who follow the Strawsonian compatibilist tradition (Strawson, 1962). Nothing I say in this section hinges on the exact argument that I discuss. However, as this is a thesis about freedom, I have chosen to discuss the moral responsibility first route as it argues to the conclusion that actions caused by non-deliberative causes are unfree.

[64]I am assuming that something like the [Principle of intercession] is a necessary condition for moral responsibility. The [Principle of intercession] can be thought of as a corollary to some readings of the principle of alternative possibilities (Frankfurt, 1969) and the principle of avoidable blame (Otsuka, 1998). Simpler accounts, such as moral influencing accounts, hold that counting people as morally responsible is justified when it influences or pressures

4) [No intercession] If *A* is caused by a non-deliberative cause, *S* cannot intercede on *A*.
5) [Lemma A] *S* cannot intercede on *A* (from 2 and 4).
6) [Lemma B] *S* is not morally responsible for *A* (from 3 and 5).
7) [Conclusion] Therefore, *A* is unfree (from 1 and 6).

According to this argument, the reason that actions caused by non-deliberative causes are unfree is because we cannot be morally responsible for them. The reason we cannot be morally responsible for actions caused by non-deliberative causes is because we are unable to intercede on actions caused by non-deliberative causes. The argument is valid.

The premise I will be challenging in this section is [No intercession]. I will propose that while someone cannot intercede on actions caused by a non-deliberative cause *at the time of action,* they were able to intercede, in a more general sense, at the time of creating the non-deliberative cause. Here's the basic idea: when someone performs the action of creating a non-deliberative cause, they are also performing a number of temporally extended actions as well.[65] After all, when someone creates a non-deliberative cause, they do so knowing what the likely future outcomes will be and they causally impact their future psychology in order to ensure those outcomes will occur. We are morally responsible for actions caused by a non-deliberative cause because we were able to intercede at the time when we created the non-deliberative cause. When we create a non-deliberative that we thought was best, we could have created a different non-deliberative cause or no non-deliberative cause at all.[66] If this proposal is right, then we are morally responsible for the consequences of our non-deliberative causes and with respect to the second objection we no longer have any reason to think the actions they cause are unfree.[67]After

---

people to behave in morally desirable ways (see for example, Schlick, 1935; Smart, 1961; though see also Arneson, 2003, Vargas, 2013). However, moral influencing accounts are contentious and most theorists do not think that pressuring and influencing people into behaving in morally desirable ways is a sufficient justification for our practices surrounding moral responsibility.

[65]This proposal has been adapted from Braddon-Mitchell (forthcoming) *Freedom and Binding Consequentialism*. This paper centres on *strong* moral dispositions that are in some ways just like non-deliberative causes. Someone creates a strong moral disposition to ensure that their future choice options are causally constrained. While there is no description of strong moral dispositions that cause actions directly, I see no reason why there could not be cases where it's our future actions, not just our future choice options, which are causally constrained.

[66]I discuss whether foreseeability is required for you to be responsible for extended actions in §5.3.1.

[67]Showing we are morally responsible for the consequences of our non-deliberative causes does not hinge on facts around action individuation, which is a contentious literature (for seminal papers see Ginet, 1990; Goldman, 1970; Davidson, 1971). There are other routes that I could have taken that would give us the same result. For example, I could have used a general moral tracing principle. We can be morally responsible for actions not under our control provided they can be traced back to actions that were under our control. While it is not possible to intercede on the actions that are caused by a non-deliberative cause, it was possible intercede on the non-deliberative cause, which caused the action, when it was being created. As a result, we are morally responsible for the consequences of our non-deliberative causes in virtue of being morally responsible for the creation of our non-deliberative causes. One way to understand moral tracing principles is as accepting [No intervention] but rejecting [Principle of Intervention]. What it hangs on is whether you think interceding is only at the time of action, or whether it is more general than this. For an in depth discussion about moral tracing approaches and their limitations see Fischer and Tognazzini (2009; 2012), McKenna (2005; 2008) and Vargas (2005).

I have finished developing this proposal, I will then briefly outline in §5.3.2 how it interacts with non-deliberative causes that don't directly cause actions but instead influence other action procedures.

**5.3.1 Moral responsibility and actions caused by non-deliberative causes**

For now, let's begin with a couple of cases that do not involve a non-deliberative cause. Imagine someone named Matt, who presses a button knowing that it will result in a nuclear bomb being dropped on Auckland, New Zealand, in twelve months. Matt knows that once the button has been pressed there is no way to intercede to prevent the nuclear attack. Matt stands before the panel and after deliberating over the various options that are available to him freely chooses to press the button. As it so happens, twelve month pass and the nuclear bomb attack occurs over Auckland, New Zealand. I don't think that it should be contentious that Matt is morally responsible for the nuclear attack on Auckland. Of course, Matt was not able to intercede at the time of the nuclear attack in order to prevent it (or for that matter, any time after pressing the button), however he was able to intercede, in some more general sense, by refraining from pushing the button. I will return to this point shortly. Importantly, if you think that Matt is morally responsible for the nuclear attack, then I think you are part of the way to also thinking that people are morally responsible for actions caused by non-deliberative causes.

Let's put the case of Matt to one side for the moment. Non-deliberative causes are much more complex than actions with foreseeable consequences because they can cause a number of different probabilistic or otherwise unforeseeable outcomes. Imagine, now, someone named Katie who, while standing next to her friend Percy, pulls a lever. The lever is part of a device that contains a complex algorithm that is capable of causing almost every conceivable outcome in the world. Katie knows about the algorithm and knows generally what sorts of factors makes the device cause good outcomes and what sorts of factors make the device to cause bad outcomes. Still, Katie is ignorant of exactly how the algorithm works and all the significant factors that determine the outcome the device will cause.[68] As it so happens, Katie pulls the lever because she thinks it would be best and as a result the device kills Percy, her friend standing next to her. I think that Katie is morally responsible for the death of Percy. Once again, I don't think that it should be contentious that Katie is morally responsible for the death of Percy. Importantly, if you think that Katie is morally responsible for the death of Percy, then I think you are part of the way to also thinking that people are morally responsible for the actions caused by non-deliberative causes. Shortly, I will draw upon the cases of Matt and Katie in order to give an account of moral responsibility for actions caused by non-deliberative causes.

[68]Alternatively, the algorithm in this case could be indeterministic. Then it would not matter if Katie had complete knowledge of the algorithm and factors that influence what outcome the device causes.

*5.3.1.1. Moral responsibility and temporally extended actions.* Before returning to a case of a non-deliberative cause, getting clear on one reason we correctly count Matt as being morally responsible for the nuclear attack on Auckland will be a helpful exercise. After all, Matt was not able to intercede on the nuclear attack at the time that the attack was taking place, and with respect to the second objection this *prima facie* suggests he is not morally responsible for the nuclear attack. However, I have also stated that Matt could have interceded, in some more general sense, how should we make sense of this claim? On one way of individuating actions, when Matt performs the action of pushing the button he also performs the temporally extended action of attacking Auckland with a nuclear weapon. Why should we think that? Well, pushing the button causally impacts the future and ensures that a nuclear attack will occur twelve months into the future. When Matt performs the action of pushing the button he does so knowing what the future outcomes of him pushing the button will be. That is, he knows that pushing the button will cause a nuclear attack to occur in Auckland in twelve months and that he will be unable to intercede and stop the nuclear attack from occurring. The upshot of this proposal is that it is clear that Matt could have interceded on the nuclear attack on Auckland. After all, he could have interceded right at the beginning of the case and refrained from pushing the button. Importantly, the fact that Matt could have interceded on this temporally extended action gives us one reason to think we are correct to count Matt as being morally responsible for the nuclear attack.[69]

*5.3.1.2. Moral responsibility and non-deliberative causes.* So far I have presented two different cases for two different purposes. The case of Matt showed that we are morally responsible for temporally extended actions and the case of Katie showed that we are be morally responsible for probabilistic outcomes. If you think that this is right then I have what is required to show that we are morally responsible for the actions caused by our non-deliberative causes. To illustrate this claim, let me bring back a case involving a non-deliberative cause. Recall the case of Bud that I described in detail earlier in this chapter. Bud created a non-deliberative cause that causes him to physically retaliate against anyone who attacks him. Imagine now that twelve months after Bud created the non-deliberative cause a frail man named Milhouse attacks Bud. Bud is not threatened by the attack and if the option were available Bud could have deescalated the situation

[69]As I stated in Footnote 67 this is not the only proposal that I could have developed to get this result. I could instead accept a stronger version of intercession such that that it does not count as ability to intercede if the possible intercession is not at the time of action. But with this stronger version of intercession, [Principle of intercession] becomes less plausible, and I would argue against it. This approach would be similar to what some have called a 'moral tracing principle'. While it is not possible for Matt intercede on nuclear attack on Auckland when it was happening, it was possible for Matt to intercede on the button press, which caused the nuclear attack on Auckland. As a result, Matt is morally responsible for the nuclear attack on Auckland in virtue of being morally responsible for the button press, which caused the nuclear attack on Auckland.

with no risk to himself or others. Unfortunately, Bud's non-deliberative cause causes him to physically attack Milhouse, Bud punches Milhouse, causing his death.

The second objection suggests that Bud is not free when he throws the punch against Millhouse because he is not morally responsible for that punch. And the reason Bud is not morally responsible for the punch that kills Milhouse is because Bud is unable to intercede on actions caused by his non-deliberative cause. Still, I think that many people (including myself) think that Bud is morally responsible for the death of Milhouse. By combining lessons from the case of Matt and Katie I can now give an account of why we correctly count Bud as being morally responsible for Milhouse's death.

On one way of individuating actions, when Bud performs the action of creating his non-deliberative cause he also performs a number of temporally extended actions as well. After all, when we create a non-deliberative cause we causally impact our future psychology in order to ensure that certain actions occur in the future. When Bud performs the action of creating his non-deliberative cause he does so knowing the kinds of future actions his non-deliberative cause will cause. That is, he knows that creating the non-deliberative cause will cause him in the future to physically retaliate against anyone who attacks him and he knows he will be unable to intercede to stop himself from doing so. This was the lesson from the case of Matt.

Let's build on this. When Bud is performing the action of creating his non-deliberative cause, and alongside that a number of temporally extended actions, he cannot possibly know the exact outcomes of his non-deliberative cause. Bud might know generally what kinds of factors will result in his non-deliberative cause causing good outcomes and bad outcomes respectively. Further, he might know generally whether the non-deliberative cause will result in more good than bad outcomes overall. For example, if the person who attacks Bud is harder, better, faster, stronger, then the act of physically retaliating might result in a bad outcome. Similarly, if the person who attacks Bud is non-threatening and frail, such in the case of Milhouse, then the outcome will be bad as well. The exact outcome of each temporally extended action is probabilistic, as Bud cannot foresee exactly which factors accompany each temporally extended act and whether they result in the outcome being good or bad each time. However, just because the outcome of each temporally extended action is probabilistic each time doesn't mean Bud is not morally responsible for them. This was the lesson from the case of Katie.

In summary, when Matt performs the action of pressing the button he also performs the temporally extended action of a nuclear attack against Auckland. That's because when Matt pushes the button he causally impacts the future to ensure a nuclear attack will occur. Similarly, when Bud performs the action of creating a non-deliberative cause he is also performing the

temporally extended action of punching. When Katie pulls the lever of the device she is morally responsible for the probabilistic outcome of Percy's death. Similarly, when Bud performs the action of creating his non-deliberative cause he is morally responsible for the probabilistic outcome of killing Milhouse. By combining these cases we then have an account of why we correctly judge someone to be morally responsible for the consequences of their non-deliberative causes. To the best of my knowledge there is no morally significant difference between these kinds of cases, unless you think something important hinges on the difference between pushing buttons, switching levers, and creating a non-deliberative cause. But this difference doesn't strike me as important with respect to determining whether or not someone counts as being morally responsible.

Importantly for my purposes, because Bud can intercede on the temporally extended action of punching Milhouse, either by not creating the non-deliberative cause or creating a different non-deliberative cause, he counts as being morally responsible for causing Milhouse's death. Further, because Bud is morally responsible for the consequences of his non-deliberative cause, relative to the second objection, we no longer to have a reason to think that Bud is acting unfreely when he punches Milhouse.

**5.3.2 Moral responsibility under the influence of non-deliberative causes**

There is something more to say about the relationship between non-deliberative causes and moral responsibility. Non-deliberative causes don't just cause actions but can also influence our other action procedures. Sometimes this influence can restrict what options are open for someone to freely perform. Earlier in this chapter I described in some detail the case of Bodi. Bodi wanted to perform the action of killing in order to prevent a truck driver from burning to death. Unfortunately, Bodi possessed a non-deliberative cause that prevented him from killing people, so killing the truck driver was not an option that Bodi could freely choose to perform. With that said, Bodi was still free to choose to perform a number of other options. Call these his **choice options**.

While we might want to count Bodi as being either morally responsible for having done the right thing or wrong thing overall, *prima facie* it appears as though Bodi is in fact morally responsible for two different actions. First, Bodi is morally responsible with respect to the action of creating the non-deliberative cause that causally constrains his future choice options. Second, Bodi is morally responsible with respect to which of his remaining choice options he freely chooses to perform. Bodi can count as being morally responsible for having done the right thing or the wrong thing with respect to each of these actions. The reason this complicates matters is because it's possible for our right/wrong judgments to be incongruent across the two actions.

For example, someone can do the right thing with respect to their action of creating a non-deliberative cause, yet do the wrong thing with respect to which of their remaining choice options they freely choose to perform.

Following Braddon-Mitchell (forthcoming) I think that whether or not someone counts as being morally responsible for having done the right thing or wrong thing overall depends on which of their actions we are focusing on and seeking to morally influence. I will not spend much time on this issue as it's unrelated to the second objection and doesn't speak directly to the issues of this chapter, which is whether or not the actions caused by non-deliberative causes are free and whether their influence is freedom undermining. However, at the very least it is of interest to very briefly sketch each possible right/wrong combination and to see whether we are likely to think that someone is morally responsible for having done the right thing or wrong thing overall. For ease of explication, I will stick with the case of Bodi.[70]

*5.3.2.1 The right non-deliberative cause and the right remaining option.* This case is simply the original Bodi case presented earlier in this chapter. Bodi does the right thing when he performs the action of creating the non-deliberative cause that removes the option of killing (see §5.2.2). While Bodi cannot shoot the truck driver he freely chooses to perform the best available remaining option which was to stay with the truck driver and comfort him the best that he can. Given the fact Bodi does the right thing with respect to both actions in this case we should have no issue judging that Bodi is morally responsible for having done the right thing overall.

*5.3.2.2 The wrong non-deliberative cause and the wrong remaining option.* Imagine, now, Bodi* who is a cunning serial killer who is deciding how he should kill his next victim. Bodi* freely created a repugnant non-deliberative cause that restricts his choice options to either killing his victim slowly and painfully, or killing his victim quickly and painlessly. The victim pleads with Bodi* to let them go but this is not a choice option because of the non-deliberative cause. As it turns out, Bodi* freely chooses to perform the worst available option and kill his unfortunate victim slowly and painfully. It should be uncontentious that Bodi* does the wrong thing when he performs the action of creating the repugnant non-deliberative cause. Similarly, it should be uncontentious that Bodi* does the wrong thing when he freely chooses to perform the worst remaining available option, killing his victim slowly and painfully. Given the fact Bodi* does the wrong thing with respect to both actions in this case we should have no issue judging that Bodi* is morally responsible for having done the wrong thing overall.

*5.3.2.3 The right non-deliberative cause and the wrong remaining option.* Now back to Bodi. Bodi plausibly does the right thing when he performs the action of creating the non-deliberative cause that

[70]This proposal has been adapted from Braddon-Mitchell (forthcoming) *Freedom and Binding Consequentialism.*

removes the option of killing. But now imagine that instead of choosing to comfort the truck driver as best he can, Bodi runs away from the scene leaving the truck driver to die alone. In this case it seems as though Bodi chooses the wrong option. While Bodi does the right thing with respect to the creation of the non-deliberative cause, he does the wrong thing with respect to which choice option he freely chooses to perform. I think in this case many people (including myself) focus on the fact that Bodi chooses the wrong choice option. As a result, we judge that Bodi is morally responsible for having done the wrong thing overall.

*5.3.2.4 The wrong non-deliberative cause and the right remaining option*. Finally, Bodi* does the wrong thing when he performs the action of creating his repugnant non-deliberative cause. However, now imagine when the victim pleads with Bodi* to let them go, and Bodi* freely chooses to do the right thing by his lights and kill them quickly and painlessly. Remember that the only other choice option available to Bodi* is to kill his victim slowly and painfully. While Bodi* does the wrong thing with respect to the action of creating the repugnant action procedure, he seemingly does the right thing with respect to which choice option he freely chooses to perform. I think in this case many people (including myself) focus on the fact that Bodi* does the wrong thing when he performs the action of creating the repugnant non-deliberative cause. The reason Bodi*'s act of murdering someone quickly and painlessly turns out to be right *simpliciter* is because his repugnant non-deliberative cause causally constrained his future choice options to just those with murderous outcomes. As a result, we judge that Bodi* is morally responsible for having done the wrong thing overall.

**5.3.3 Conclusion: Non-Deliberative Causes and Moral Responsibility**

In this section I have discussed the charge that because we are unable to intercede on actions caused by non-deliberative causes, we cannot be morally responsible for them. Importantly, because there is a tight conceptual connection between moral responsibility and acting freely, if we are not morally responsible for the actions caused by our non-deliberative causes, then they are unfree. In reply, I have given one account (but it's by no means the only possible account) of why we correctly count people as being morally responsible for actions caused by non-deliberative causes. This is because people *are* able to intercede upon actions caused by non-deliberative causes. This account drew on the cases of Matt and Katie that showed that we are morally responsible for temporally extended actions and probabilistic outcomes respectively. By combining these two cases we get an account of why we count as morally responsible for actions caused by our non-deliberative causes. For example, when Bud performed the action of creating his non-deliberative cause, he also performs a number temporally extended actions including the critical action of punching Milhouse. Further, even though the outcomes of these temporally

extended actions are probabilistic, such as the death of Milhouse, Bud is still morally responsible for them. That's because when Bud creates the non-deliberative cause he knows generally what actions it will cause and causally impacts his future psychology in order to ensure those actions occur. Further, Bud generally knows the kinds of factors that will result in his non-deliberative cause causing good and bad outcomes and whether or not the non-deliberative cause will result in more good or bad outcomes overall.

The main upshot of this proposal is that because the temporally extended actions occur alongside the creation of the non-deliberative cause it makes sense that someone could intercede on these temporally extended actions. Specifically, they could intercede by not creating the non-deliberative cause, or by creating some other non-deliberative cause. This means that people are morally responsible for the consequences of their non-deliberative causes, and importantly shows that we no longer have a reason, with respect to the second objection, to think that actions caused by non-deliberative causes are unfree.

### 5.4. Objection 3: Indirect compatibilism, degrees of freedom and vagueness[71]

Before proceeding, it will be helpful once again to briefly remind ourselves of some of the key definitions and components of indirect (conscious control) compatibilism. According to indirect compatibilism someone performs an action freely iff the action is caused by freedom-conferring mechanisms (FCM), or controlled action procedures. Actions count as directly free if they are caused by a freedom-conferring mechanism, and count as indirectly free if they are caused by a controlled action procedure. Controlled action procedures come in two varieties: paradigmatic and non-paradigmatic. Something is a paradigmatic controlled action procedure (P-CAP) iff it's a controlled action procedure that fulfils all the direct control conditions. This means that it is created, and could be destroyed, triggered, monitored, and updated by a freedom-conferring mechanism or another controlled action procedure. Actions caused by P-CAPs are paradigmatically indirectly free. Something is a non-paradigmatic controlled action procedure (NP-CAP) iff it's a controlled action procedure that does not fulfil at least one of the direct control conditions. This means that it was either not created, could not be destroyed, triggered, monitored, or updated by a freedom-conferring mechanism or another controlled action procedure. Actions caused by NP-CAPs are non-paradigmatically indirectly free.

According to the third objection there seems to be something *prima facie* wrong with indirect compatibilism if it counts directly free actions, paradigmatically free actions and non-paradigmatically free actions as all being fully free. After all, there seems to be something very

[71]Thanks and credit to Kristie Miller for her assistance with this objection and section.

different between cases involving FCMs, P-CAPs and NP-CAPs. We are inclined to say, straightforwardly, that actions caused by FCMs are free. However, whether or not an action caused by a controlled action procedure is free seems to be a graded matter that depends on the number of direct control conditions – creation, destruction, triggering, monitoring, and update – that are fulfilled and the extent to which they are fulfilled. Direct control conditions are graded from having none at all to any extent, to having all of them to the full extent. [72] I will return to discuss exactly what this amounts to shortly. For now, it will be natural to suppose that we can model the degree to which an action is free, by supposing that actions that are free to degree 0 are not free at all, and those that are free to degree 1 are fully free. The remainder are free to some degree between 0 and 1.

For ease of explication I am going to assume that the degree to which an action is free is the average degree to which the action procedure which causes the action fulfils the direct control conditions. So for example, a P-CAP with creation to degree 1, destruction to degree 1, triggering to degree 1, monitoring to degree 1, and update to degree 1 has direct control conditions to degree 1. This means that the actions this P-CAP causes are fully free. Before now, I have supposed that P-CAPs were action procedures that fulfilled all 5 direct control conditions, but from now on I will say that they are action procedures that fulfil all 5 conditions *to degree 1*. This means that there can be NP-CAPs that fulfil all 5 direct conditions, but do so to less than degree 1. So for example, an NP-CAP with creation to degree .75, destruction to degree .25, triggering to degree .25, monitoring to degree .50, and updating to degree .75 has direct control conditions to degree .5 (the average of all of these). This means that the actions this NP-CAP causes are free to degree .5. Finally, an NP-CAP with creation to degree 1, destruction to degree 0, triggering to degree .25, monitoring to degree .50, and updating to degree .75 has direct control conditions to degree .5. This means that the actions this NP-CAP causes are also free to degree .5.

I think we are inclined to count actions caused by P-CAPs as free because they have all 5 of direct control conditions and those conditions are fulfilled to the full extent. Further, the further away an NP-CAP gets from a P-CAP, the less I think we are sure that the action it causes is free. That's because the further an NP-CAP is away from a P-CAP the lesser the degree of direct control conditions it has, either because it lacks some direct control conditions, or it only fulfils its direct control conditions to a limited extent, or some combination of both. What, then,

[72]There are possible complications involving the ranking, trumping, and commensurability between the various kinds of direct control conditions. Unfortunately, these are not issues I am able to address in this thesis. For the purposes of this section I don't need the correct method of determining the overall degree of direct control conditions, instead what I need are accounts of how freedom of action depends on the degree of direct control conditions, however it is that is best calculated.

is required to address the third objection is an account of how freedom of action depends on the degree of direct control conditions of the action procedure which causes the action.

In what follows I will present three broad kinds of accounts about how the freedom of an action depends on the degree of direct control conditions of the action procedure which causes the action. Remember that the purpose of doing this is to show how we might accommodate the range of intuitions that motivate the third objection, and to do so in a manner that is compatible with indirect compatibilism. Which account you think is best will depend on your various intuitions about free actions, and for my purposes, I don't need to take a stand on which is preferable. In §5.4.1 I will begin by outlining some accounts that are consistent with the orthodox intuition that a necessary condition for an account of free action is that being free is all or nothing. Then, in §5.4.2 I will outline some accounts that are consistent with the intuition that acting freely can come in degrees, and that its possible for actions to be more or less free. Finally, in §5.4.3 I will outline an account that is consistent with thinking acting freely comes in degrees and that only actions caused by FCMs are fully free.

### 5.4.1 Freedom of action, direct control conditions and vagueness

Let's begin with the simplest account of the relationship between freedom of action and degrees of direct control conditions. For any action *A*:

> **[Basic account]:** *A* is fully free if the action procedure that caused *A* has *any* degree of direct control conditions.

Up to this point in the thesis I have assumed something like [Basic account] for ease of explication. Obviously this proposal will not do, as it is unable to accommodate the intuitions that motivate the third objection. The reason that we are inclined to distinguish actions caused by FCMs, P-CAPs and NP-CAPs is because our intuitions about cases of each kind are rather different. After all, it seems manifestly wrong that an action caused by a CAP with direct control conditions to degree 0.1 and another action caused by a CAP with direct control conditions to degree 0.9 are both fully free. However, insisting that actions caused by FCMs, P-CAPs and NP-CAPs are all fully free is not something that indirect compatibilism need be committed to. In what follows, I will outline some different proposals regarding the relationship between our freedom of action and the degrees of direct control conditions that can accommodate the intuitions behind the third objection.

The major problem with the basic proposal is that it seems manifestly false that actions caused by NP-CAPs with small degrees of direct control conditions should count as fully free. Suppose we accept this intuition, alongside the intuition that being free is an all or nothing matter: it does not come in degrees. One way to accommodate these two thoughts is by

introducing some straightforward threshold. In that case, the freedom of an action is not guaranteed by the CAP having a non-zero degree of direct control conditions; instead it requires the CAP to have direct control conditions above a certain degree. So, for any action *A*:

> **[Straightforward threshold]**: *A* is fully free if the action procedure that caused *A* has some degree of direct control conditions either equal to or above a *straightforward* threshold.

If [Straightforward threshold] is right, then there is a straightforward threshold that cleaves the actions that count as fully free from those that are not free at all. For the sake of illustration let's imagine that the straightforward threshold value is 0.5. Actions caused by CAPs with direct control conditions either equal to or above degree 0.5 are fully free. Conversely, actions caused by CAPs with direct control conditions below degree 0.5 are unfree. Thus actions are either free, or not free, rather than free to some degree between 1 and 0.

It's interesting to note that if [Straightforward threshold] is right, then it might be possible to empirically discover what the straightforward threshold value of direct control conditions required to be fully free is. While I think that this is a worthwhile project to attempt in the future, I think the prospects of such a project succeeding are limited, for reasons I will outline below.

There are two major issues with [Straightforward threshold]. First, I think it's unlikely there is a straightforward threshold value that would successfully cleave fully free actions from unfree actions. For example, I might perform an empirical study and find that one group of people set the straightforward threshold at 0.4 and another group sets the straightforward threshold at 0.6. Perhaps the people who are setting the straightforward threshold at 0.4 are doing so to accommodate cases that they consider to be important cases of free action. Perhaps at least one of these groups is wrong, because there is a metaphysical fact about the correct straightforward threshold value, a fact that is, for instance, generated by people's dispositions. But I am not confident that this is so. All that's important for my purposes, however, is that there is unlikely to be any convergence around a single straightforward threshold value (even if there is in fact a correct straightforward threshold value).

Second, I think the idea there is a sharp cut off between which actions count as fully free, and which as unfree, seems wrong. Once again, imagine that the straightforward threshold value is 0.5. Actions caused by CAPs with direct control conditions to degree 0.5 would be fully free. Conversely, actions caused by CAPs with direct control conditions to degree 0.49 would be unfree. Further, those latter actions would be just as unfree as actions caused by action procedures that have no direct control conditions. This just seems wrong. Cases involving direct

control conditions to degree 0.49 seem to us to be closer to cases involving direct control conditions to degree 0.5, and on this view those cases count as fully free. However, given the straightforward threshold value of 0.5, cases involving direct control conditions to degree 0.49 are just as unfree as cases involving direct control conditions to degree 0. I think what these two issues suggest is that whether or not some actions count as fully free or unfree is, in some sense, a vague matter.

One way to accommodate this thought is to think that while there is a threshold between what actions count as fully free and what actions do not, that threshold is vague and not a straightforward value. That is, whether or not an action counts as fully free doesn't depend on the CAP having direct control conditions above some straightforward threshold value. Instead, it depends on the CAP having direct control conditions above some vague threshold. So, for any action *A*:

> **[Vague threshold]**: *A* is fully free if the action procedure that caused *A* has some degree of direct control conditions above a *vague* threshold.

Here is the basic idea: while there are clear cases of actions being fully free (those caused by FCMs and P-CAPs) and clear cases of actions being fully unfree (those caused by action procedures with no direct control conditions) there are a whole range of cases in between (those caused by NP-CAPs). To allow for vagueness, then, would be to allow that the expression 'free action' admits of a number of borderline cases: actions that are neither determinately fully free, nor determinately unfree. These actions fall within the *penumbra*.

Some (or perhaps even all) of the cases that fall within the penumbra will be those involving NP-CAPs, actions that I have said count as non-paradigmatically indirectly free. So rather than simply distinguishing directly free actions, and indirectly free actions that are paradigmatic and non-paradigmatic, we can instead suggest that there is a range of actions for which it is *vague* whether or not those actions count as free. These are the actions that fall within the penumbra. We can further accommodate many of our intuitions in these cases by noting that some of these actions are more towards the centre of the penumbra and some are more towards one of the edges of the penumbra. That is, some actions within the penumbra are closer to being determinately free and some are closer to being determinately unfree. One upshot of this view is we can accommodate the idea that in some sense freedom of action comes in gradations without actually needing to posit that freedom of action itself comes in degrees. Further, if one was worried that there would be an objectionably sharp cut-off between cases of an action being determinately free (or unfree) and being in the penumbra, we can posit higher-order vagueness. If there is higher-order vagueness, then the very borders of the penumbra are themselves vague.

As a result, it is an indeterminate matter whether, for some actions, it is indeterminate whether those actions are fully free. Once again, the fact that there can be determinate penumbral cases of free action and indeterminate penumbral cases of free action can explain a great many of our intuitions about cases I have said count as non-paradigmatically indirectly free.

Once again, for my purposes it doesn't matter which account of vagueness you prefer. It could be that vagueness is purely epistemic.[73] That is, there could be a fact of the matter regarding which actions count as fully free, and which actions count as unfree, with nothing in between. It's worthwhile to note that [Straightforward threshold] is compatible with epistemic vagueness. On this view, there aren't really any penumbral cases of free action. Instead, the penumbra simply reflects the fact that we don't know which actions are determinately fully free, and which actions are determinately unfree. The reason we don't know this is because we don't know precisely what the extension of the expression 'free action' is, and what's more, we cannot know this. Still, the expression 'free action' does have a precise extension and so there is a fact of the matter, for each action, whether or not that action counts as fully free. Nevertheless, since we can't know where the cut-of is, between which actions count as fully free, and which actions count as unfree, this view will still explain our intuitions on the matter.

By contrast, it might be that vagueness issues entirely from our language and concepts. As the expression 'free action' is imprecise in various ways, given what we mean by said expression, for some actions it is simply indeterminate whether or not they count as fully free. On such a view our concept of free action is vague, so that it simply admits of a whole range of *possible* precisifications, with each precisification being just as good as the other. For each precisification there is a clear answer to the question of whether or not the action counts as fully free or as unfree, but there is no reason to prefer one precisification over the others. As a result, it is a vague matter, for some actions, whether or not they count as fully free, or as unfree. Those who are attracted to supervaluationist[74] or subvaluationist[75] accounts of vagueness will likely think this is right. The difference between these two views is that supervaluationists think that there are truth-value *gaps*. They suppose that the actions that fall within the penumbra are such that it is neither true, nor false, that the action counts as fully free (or unfree). In contrast, subvaluationists think that the actions that fall within the penumbra are such that it is *both* true

[73]Those who defend epistemic approaches to vagueness include Sorensen (1988; 2001), Williamson (1994; 2000), Graff (2000), Fara (2008), and Rescher (2009).
[74]Those who have defended supervaluationist accounts of vagueness include Fine (1975), Keefe (2000), Lewis (1986), and Varzi (2001).
[75]Those who have defended subvaluationist accounts of vagueness include Corberos (2011), and Hyde and Colyvan (2008).

and not true, that the action counts as fully free. According to the subvaluationist there are no truth-value gaps, instead there are truth-value *gluts*.

I am not going to take a stand on which of these approaches is superior, as this will depend on a whole range of other background beliefs about which view of vagueness is superior. All that matters for my purposes is that each of these approaches is consistent with indirect compatibilism, and that each can accommodate the range of intuitions that motivate the third objection.[76]

### 5.4.2 Degrees of freedom and direct control conditions

So far I have only provided accounts that are consistent with the orthodox intuition that a necessary condition for an account of free action is that being free is an all or nothing matter. However, I think it is very tempting to say that any distinction in freedom between actions caused by FCMs, P-CAPs and NP-CAPs is a matter of degree rather than being an all or nothing matter. [77] If direct control conditions are graded and come in degrees then why think that it's an all or nothing matter whether or not an action is fully free. Instead, our actions can be more, or less, free: that is, it is a matter of degree how free they are. [78] One basic way to account for degrees of freedom is with a straightforward mapping between degrees of freedom and the degrees of direct control conditions. So, for any action *A*:

> **[Degrees of freedom]** *A* is free to a non-zero degree if the action procedure that caused *A* has *any* degree of direct control conditions.

The most natural way to think of indirect compatibilism in terms of [Degrees of freedom] is to think that directly free actions (those caused by FCMs) and actions caused by P-CAPs are fully free and that actions caused by NP-CAPs are free to some degree less than 1. Specifically, actions caused by an action procedure with no direct control conditions do not count as free to any degree. Actions that are caused by NP-CAPs with direct control conditions to some in-between degree, are free to some in-between degree. Finally, actions that are caused by P-CAPs

[76]It if it is a vague matter, for some actions, whether those actions are free, then by extension it can also be a vague matter whether there is free will. Recall that on my account whether or not we have free will is determined by whether a sufficient percentage of our actions are free. Of course, if we are epistemicists about the vagueness of free action then we will think that if it is vague whether or not we have free will this is because although there is some determinate fact of the matter we cannot know this fact. In contrast, supervaluationists will think that it is simply semantically underdetermined whether or not we have free will. On some precisifications we do and on others we do not, hence it is vague whether or not we do.

[77]Current orthodoxy in the literature counts necessary condition for our freedom is that it is all or nothing. I won't attempt to argue this current orthodoxy is wrong in this chapter. Instead, I will just outline how an account of freedom coming in degrees that naturally falls out of indirect compatibilism. To the best of my knowledge the only other account of degrees of freedom in the literature is due to O'Connor (2005; 2009).

[78]Once again, if it can be a matter of degree whether or not our actions are free, then by extension, so too can it be a matter of degree whether or not we have free will.

with direct control conditions to degree 1, and FCMs, are fully free.[79] I will have something to say to those who might want a way to distinguish P-CAPs and FCMs, but for now just follow me in thinking the actions in both cases are fully free. One benefit of this view is that it is flexible in accommodating a range of intuitions about free actions. We don't just have to talk about directly free actions and indirectly free actions that are paradigmatic and non-paradigmatic. Instead, within non-paradigmatically indirectly free actions we can distinguish a whole range of gradations in the extent to which those actions are free.[80]

While I think the idea that freedom of action comes in degrees is probably right, [Degrees of freedom] is obviously wrong. That is because it's very unlikely that there is a straightforward 1-to-1 mapping between the degrees of freedom of an action and the degrees of direct control conditions a CAP has. For example, consider an action caused by an NP-CAP with direct control conditions to degree .001. It seems to me that we don't think that such an action is free to degree .001; instead we would think, straightforwardly, that such an action is just unfree. Again, one way to accommodate this thought is to think that there is a vague threshold between what actions count as free to some non-zero degree, and what actions are unfree (see §5.4.1). Whether or not an action counts as being free to a non-zero degree depends on the CAP

[79]Someone might think that the only actions that count as fully free are those caused by FCMs. Actions caused by P-CAPs might be free to a high degree, but they are fully free. I will describe such a view shortly.

[80]There are a number of different ways we can think about the degree theoretic nature of free action. The idea that there are degrees of truth is formulated within a many-valued logic. Those who have defend such logics include Dunn and Hardegree (2001), Shramko and Wansing (2011), Gottwald (2001), and Weber and Colyvan (2010). Not all many valued logics have infinite values, some have as few as three (see for example Tye, 1994). For my purposes, it really doesn't matter what account you prefer. One could think what comes in degrees is not the direct control conditions, but instead, truth. That is, propositions of the form [x is a free action] are (when appropriately filled out) true to a degree between 0 and 1. If they are true to degree 0 then they are not true, and if they true to degree 1 then they are maximally true, and truth in between, is just that, truth in between. The proposition [x is a free action] will be free to degree 1 for directly free actions (those caused by FCMs) and true to some degree for indirectly free actions (those caused by P-CAPs and NP-CAPs). Specifically, the proposition [x is a free action] will be free to degree 0 for actions caused by an action procedure with no direct control conditions. The proposition [x is a free action] will be free to a degree in between for actions caused by NP-CAPs, which have some amount of direct control conditions. Finally, the proposition [x is a free action] will be free to a maximal or high degree for actions caused by P-CAPs, which have the maximal amount of direct control conditions.

Of course someone doesn't have to understand this view in terms of degrees of truth. Instead of truth coming in degrees it could be that the world itself has a degree theoretic structure. As a result, it could be that direct control conditions can be instantiated in degrees and an action can instantiate the property freedom to a degree between 0 and 1. The idea that there is a degree-theoretic structure to the world arose out of the idea of fuzzy logics (Zadeh, 1965). Those who defend such approaches include Hyde (2008) and Smith (2008). If an action procedure instantiates the direct control conditions to degree 0, then the action it causes is unfree, and if it instantiates to degree 1, then it is free to a maximal or high degree, and if it instantiates it to a degree in between 1 and 0, then it is free to some in between degree. Directly free actions (those caused by FCMs) are fully free and indirectly free actions (those caused by P-CAPs and NP-CAPs) are free to some degree. Specifically, actions that are caused by P-CAPs, which have the maximal amount of direct control conditions, are free to a maximal or high degree. Further, actions that are caused by NP-CAPs, which possess some amount of direct control conditions, are free to some in between degree. Understood in this way, claims about whether or not an action is free or not, ought properly to be framed as claims about the degree to which an action is free. These claims, moreover, will simply be true or false. For example, the proposition [x is free to degree 0.1] is true, iff x is free to degree 0.1. So truth itself does not come in degrees.

that caused it having a degree of direct control conditions above some vague threshold. So, for any action *A*:

> **[Degrees of freedom with vague threshold]** *A* is free to a non-zero degree if the action procedure that caused *A* has a degree of direct control conditions above a *vague* threshold.

If [Degrees of freedom with vague threshold] is right, then whether or not an action counts as free to a non-zero degree is not just a straightforward matter of reading the degree of direct control conditions the CAP has; at least for the lowest values of direct control conditions. However, this only seems to be part of the picture. We might also think there is a second vague threshold between those actions which count as fully free and those actions which are free to a non-zero degree. For example, consider an action caused by an NP-CAP with direct control conditions to degree .999. Again, it seems to me that we don't think that such an action is free to degree .999; instead we would think, straightforwardly, that such an action is just fully free. Once again we can accommodate this thought by thinking that there should not be one vague threshold but two: one between the actions that are fully unfree, and those that are free to some degree, and one between the actions that are fully free, and those that are free to some degree. It is worthwhile to note that any additional amount of freedom-conferring properties a CAP has above this higher threshold value (wherever that value might be) contribute nothing more to the freeness of an action on this proposal. Whether or not an action counts as being free to a non-zero degree depends on the CAP having some degree of direct control conditions above some lower vague threshold. Furthermore, whether or not an action counts as being fully free depends on the CAP having some degree of direct control conditions that is above some higher vague threshold. So, for any action *A*:

> **[Degrees of freedom with vague thresholds]** *A* is free to a non-zero degree if the action procedure that caused *A* has a degree of direct control conditions above a lower *vague* threshold and below a higher *vague* threshold. *A* is fully free if the action procedure that caused *A* has a degree of direct control conditions above a higher *vague* threshold.

Now, if [Degrees of freedom with vague thresholds] is right then, whether or not an action counts as free to a non-zero degree is not just a straightforward matter of reading the degree of direct control conditions a CAP has at the lowest values of direct control conditions. Similarly, whether or not an action counts as fully free is not just a straightforward matter of being just those cases where the CAP has all the direct control conditions. Once again, admitting of vagueness allows of a number of borderline cases: actions that are neither determinately free to any degree, nor determinately unfree. Further, there are actions that are neither determinately

fully free, nor determinately free to just a high degree. I will not reiterate my discussion of vagueness here (see §5.4.1). All that is important for my purposes here, is that by combining the resources of both degrees of freedom and vagueness, we can accommodate the full range of intuitions that are motivating the third objection in a manner that is consistent with indirect compatibilism.

### 5.4.3. Accommodating the intuition that only FCMs are fully free

There is one intuition, which I don't possess, but that I imagine some people have and is worthwhile briefly mentioning. The intuition is that actions caused by P-CAPs, despite having direct control conditions to degree 1, are never fully free: the only actions that are ever fully free are those that are caused by FCMs. Let's suppose for the moment that P-CAPs can only cause actions that are free to degree .8. This value is completely arbitrary and is only for explanatory purposes; nothing hinges on the value chosen here. Whether or not an action counts as being free to a non-zero degree depends on the CAP having some degree of direct control conditions that is above some lower vague threshold. Furthermore, whether or not an action counts as being free to a maximum degree .8 depends on the CAP having some degree of direct control conditions that is above some higher vague threshold. So, for any action *A*:

> **[Limited degrees of freedom with vague thresholds]** *A* is free to a non-zero degree if the action procedure that caused *A* has a degree of direct control conditions above a lower *vague* threshold. *A* is free to a maximum degree .8 if the action procedure that caused *A* has a degree of direct control conditions above a higher *vague* threshold.

If [Limited degrees of freedom with vague thresholds] is right, then whether or not an action counts as free to a non-zero degree is a vague matter at the lowest degree values of direct control conditions. Similarly, whether or not an action counts is free to the maximum degree of 0.8 is a vague matter at the highest degree values of direct control conditions.

Whether or not actions caused by FCMs are the only ones that count as being fully free is in part an empirical matter that I will address in a future research project. I think that actions caused by P-CAPs and even some NP-CAPs might turn out to be fully free. However, even if it turns out that I am wrong and only actions caused by FCMs are fully free, I don't see why this would turn out to be an issue for indirect compatibilism. As I hope to have shown in this section, indirect compatibilism has the tools available to accommodate a wide range of different intuitions about the status of free actions caused by FCMs, P-CAPs and NP-CAPs.

### 5.4.4 Conclusion: Indirect compatibilism, degrees of freedom and vagueness

In summary, I have addressed a third objection according to which there seems to be something wrong with indirect compatibilism if it counts directly free actions, and indirectly free actions

both paradigmatic and non-paradigmatic, as all being fully free. That's because there seems to be something very different about cases involving FCMs, P-CAPs, and NP-CAPs. In response I have outlined a number of proposals that are consistent with indirect compatibilism and can accommodate the full range of intuitions that are motivating the third objection. One way we might accommodate these intuitions is with vagueness. While actions caused by FCMs and P-CAPs are determinately free and actions caused by an action procedure with no freedom-conferring properties are determinately unfree, it is a genuinely vague matter whether or not actions caused by NP-CAPs with some in-between degree of direct control conditions are fully free or not. Another way we might accommodate these intuitions is with an account of degrees of freedom that falls naturally out of indirect compatibilism. While actions caused by FCMs and P-CAPs are fully free, actions caused by NP-CAPs are free to a degree determined by the degree of direct control conditions that they have. Most likely, in order to accommodate the full range of intuitions motivating the third objection we will need to draw upon both degrees of freedom and vagueness. Irrespective of the way we choose to do it, what is important for my purposes is the fact that we can accommodate these intuitions in a manner that is compatible with indirect compatibilism.

### 5.5. Objection 4: Getting empirical traction on indirect compatibilism

According to indirect compatibilism an action is free iff it is caused by freedom-conferring mechanisms (FCMs), or controlled action procedures (CAPs). Actions count as directly free if they are caused by an FCM, and count as indirectly free if they are caused by a CAP. According to the fourth objection everything I have said so far about my new account of free action, indirect compatibilism, has been an exercise in *a priori* reasoning. While I have extensively discussed empirical evidence from experimental philosophy and the brain sciences, I have not presented any empirical evidence that gives us empirical traction on the matter of whether or not indirect compatibilism is the correct account of free action. In response to this fourth objection I will present one new piece of evidence from my own investigations in the brain sciences that I think shows that an important empirical *necessary* condition is satisfied by my view. While this result alone does not tell us that indirect compatibilism is the correct account of free action, it does militate against some worries that the view *cannot* be true.

#### 5.5.1 Empirical evidence and indirect compatibilism

Before describing the novel empirical finding, I think it is important to get clear on the scope that empirical evidence from the brain sciences can bring to bear on indirect compatibilism. No amount of evidence from the brain sciences can show that indirect compatibilism is the right

account of free action. There is no empirical *sufficient* condition for indirect compatibilism being true. That's because whether or not indirectly free actions get to count as free is not a matter that the brain sciences can provide direct empirical traction on. There is no empirical fact of the matter here; there is simply a matter of whether the cases of indirectly free actions count, or upon reflection could, or should, come to be counted, as free. That's not to say that there is no empirical evidence that could motivate arguments for indirect compatibilism being a preferable account of free action over direct (conscious control) compatibilism. For example, future projects in experimental philosophy that I am working on might establish that the general population count indirectly free actions as free, or come to do so reliably after reflection. Such evidence would show that indirectly free actions are already counted as free by our concept of free action, or that upon reflection, in order to include these cases, we are disposed to make principled revisions to our concept of free action. Both of these results would provide some good empirical reasons to favour indirect compatibilism as an account of our free action. However, I will not pursue this any further in this thesis.

### 5.5.2 One Novel Empirical Finding from My Own Investigations

The empirical necessary condition that my finding is relevant to is whether there are in fact cases of creating, triggering and updating of action procedures. If action procedures are never created, triggered or updated by methods that depend on FCMs, then indirect compatibilism will be false for the simple reason that there are as a matter of fact no indirectly free actions. While for various reasons it is more difficult to devise protocols that test for creation and destruction,[81] I plan in future empirical work to look for evidence of triggering and monitoring, and I present in this section evidence for updating. In a paper reproduced in Appendix A (Latham et al., 2017), my co-investigators and I found electrocortical evidence of updating of CAPs in conditions which give us reason to believe that these updatings are a result of conscious control.

The key discovery of this paper, for the purposes of this thesis, was the discovery of a very late period of electrocortical activity centred over the left frontal polar cortex. This electrophysiological activity occurred after people received feedback that they had been rewarded for their experimental performance. The exact experimental conditions under which we observed this electrocortical activity will be important later, once I have provided some general details about the experimental task, as I think it gives us some further reasons to think this electrocortical activity reflects a candidate neural realizer of an FCM. For the moment I will set aside these complications and just focus on the electrocortical component itself.

[81]It's easier to devise experimental protocols that test for triggering and updating because these are processes that can be induced to occur numerous times over the course of an experimental task. It's very likely that the processes of creation and destruction only occur on a single-trial with an experimental task.

This particular electrocortical component has been observed only once before in prospective memory tasks (see for example, West, 2011; West & Moore, 2002). In prospective memory tasks people encode an intention in response to an instruction stimulus to be performed at a later time. Interestingly, in our experiment, this electrocortical component was not observed in response to an instruction stimulus. Rather, it was observed in response to a feedback stimulus that signalled reward. This is significant, as participants can't be encoding an intention to perform some specific action at a later time, as they were in prospective memory tasks, as they don't know exactly what they will be required to do next. Instead, what I think people are doing is encoding an updated CAP that will continue to automate and control their performance over the course of the experimental task. However, in order to explain why I think this I first need to outline a few general details about the experimental task.

*5.5.2.1. Behind the empirical result: Methods of experimental task.* Our experimental task involved 20 right-handed males performing a variation of the monetary incentive delay task. At the beginning of each experimental trial participants were presented with an incentive cue that indicated what the possible outcome for that trial would be: reward, loss, or neutral break-even. It also signalled what hand, left or right, participants would be required to respond with for that trial. After an anticipatory period, a target stimulus was presented that participants were required to respond to as close to 1 s after presentation as possible. If participants were successful with their performance then they were rewarded on reward trials, or avoided punishment on loss trials (henceforth I will simply refer to both these outcomes as rewards). Importantly, in our experiment the outcomes were all rigged and participants were told that successful performance would only increase the likelihood that they would receive reward. The main reason we did this was in order to increase participant engagement with the experimental task. Under these conditions of uncertainty, on any given experimental trial, participants can never be completely sure whether the reason they are being rewarded is on account of their performance or because they are lucky.

This kind of experimental task lends itself to investigating four key stages in reward and no-reward related processing. For example, we are able to examine electrocortical changes with respect to anticipatory processes during the presentation of the incentive cue. We are able to examine the modulation of target processing during the presentation of the target stimulus. We are able to examine the modulation of response preparation and responses during the simple response task. Finally, we are able to examine reward consumption processes during the presentation of the feedback stimulus. As I mentioned previously, the complete breakdown of

these various results can be found in Appendix A. What I will be focused on here is the novel electrophysiological component observed well after the feedback stimulus had been presented.

*5.5.2.2. Behind the empirical result: Interpreting the result of the experimental task.* Recall that the key discovery, for the purposes of this thesis, is a very late period of electrocortical activity located over the left frontal polar cortex. This electrocortical component has only been observed once before in prospective memory tasks, which require people to encode an intention, in response to an instruction, to be performed at a later time. Importantly, we didn't observe this component in response to any task instructions but instead in response to a feedback stimulus that signalled reward. Again, the significance of this is that people can't have been encoding an intention to perform a specific action at a later time in our experiment, as they didn't know exactly what they would be required to do next. Now that I have described some of the basic methodological details I can complicate this result slightly by noting that the key discovery was only observed after feedback stimulus that signalled reward on experimental trials performed with the right hand! I still think that what people are doing is encoding an updated CAP that will continue to automate and control their performance over the course of the experimental task. However, what could explain why people are only selectively updating their CAPs on trials where they successfully performed with their right hand?

Here is what I think is going on. Let me begin by drawing your attention to two points in the methodology that I think are important for explaining this result. First, participants were told that successful performance would only increase the likelihood that they would receive a reward, not guarantee it. This means that on any given experimental trial participants could never be certain whether or not they were being rewarded on account of their performance, or on account of luck. Second, everyone in this experiment was right handed. So, our key result was only observed when participants were rewarded after making a response with their dominant hand. People should be most confident in their performance when they are making a response with their dominant hand. As a result, on experimental trials performed with their dominant hand, people should be more inclined to think that they are being rewarded on account of their performance and not on account of luck. The reason, then, that we see participant's updating CAPs on trials performed with the dominant hand is because they are confident that their performance was causally efficacious in securing their reward. After all, the best way to maximise payoffs in the experimental task was to continually update their CAP to emulate the successful performance shown by their dominant hand. It's worthwhile to note that this explanation makes

a simple prediction for a future confirmatory study. People who are left handed should update their CAPs when rewarded following left, and not right, hand responses.[82]

### 5.5.3 Conclusion: Getting empirical traction on indirect compatibilism

In summary, I have addressed a fourth objection that is concerned I have not presented any empirical evidence that gives us empirical traction on the matter of whether or not indirect compatibilism is the correct account of free action. The only empirical evidence that directly bears on a sufficient condition as to whether or not indirectly free actions are free *simpliciter* will be from experimental philosophy, which can tell us whether or not people count cases of indirect freedom as free, or upon reflection could come to judge them as free, and this only if we hold that such experimental results can be sufficient rather than an adjunct to philosophical analysis. Still, a necessary condition for there to be indirect freedom is for there to be cases where our FCMs create, destroy, trigger, or update CAPs. Recall that in Chapter 3 and Chapter 4 of this thesis, I addressed a wide range of evidence from the brain sciences that showed that this might not occur at all. While I established that such a conclusion is untenable it left the door open whether there was in fact any positive evidence of this occurring. Here I highlight one key result from my own investigations into the brain sciences: late stage negativity over the left frontopolar cortex that occurs after feedback under conditions of uncertainty when responding with a dominant hand. If I am right, then, this is evidence that people engage FCMs even on very simple tasks, provided those tasks are incentivised and performance matters.

This evidence does bear on the question of whether or not there can be free actions at all given my view. Remember, actions are indirectly free iff they are caused by a CAP. The reason actions caused by CAPs are free is because said CAP was created by FCMs, or could be destroyed, triggered, or updated by FCMs. Therefore, it's still a necessary condition for indirect compatibilism that FCMs play the active role we suppose they do for actions caused by CAPs as well. If FCMs play no role in our action procedures, then there are no CAPs. Further, if there are no CAPs then there is no indirect freedom, so indirect compatibilism is false.

## 5.6 Conclusion: Objections and replies to indirect compatibilism

In summary, in this chapter I have worked through what I think are some of the most interesting objections to indirect compatibilism. I began this chapter by introducing non-deliberative causes.

[82]You might think that the novel electrocortical component we observe is not a candidate realizer of FCMs updating CAPs. Instead it merely realizes some unconscious, perhaps indirectly free, updating process. While that's possible, I see no reason to think that is what is occurring here. Remember, that the only other time this kind of electrocortical component has been observed is in prospective memory tasks where people consciously encode intentions, in response to instructions, to be performed in the future. Why think that what this electrocortical component realizes in our experimental task is that drastically different from what we observe in prospective memory tasks?

Non-deliberative causes are freely created action procedures that we have no ongoing control over. This means that non-deliberative causes cannot be destroyed, triggered, monitored, or updated by a FCM or other CAP. I think that non-deliberative causes are the most potent kind of putative counterexample to indirect compatibilism. So if I was able to motivate the idea that actions caused by non-deliberative causes are indirectly free then I think I can make the case that other kinds of NP-CAPs do not pose an issue for indirect compatibilism.

The first objection I addressed was that non-deliberative causes cause unfree actions because they are pathological, and the reason they are pathological is because they cause actions that appear, at the time of action, to be non-maximising. I undermined this argument in two ways. First, actions caused by non-deliberative causes that are pathological can nonetheless be free because pathological actions can be free under certain circumstances. Namely, when the non-deliberative cause does what it was intended to do. Second, some actions caused by non-deliberative causes are non-maximising at the time of action, but maximising overall. Provided you think that something that is maximising overall is not pathological then we have a good reason to think non-deliberative causes are not pathological, and hence no reason to think that they are counterexamples to indirect compatibilism.

The second objection I addressed was that because we are unable to intercede on actions caused by non-deliberative causes, we cannot be morally responsible for them. Importantly, because of the tight conceptual connection between moral responsibility and acting freely, if we are not morally responsible for the actions caused by our non-deliberative causes then they are unfree. In reply, I argued we could have interceded by not creating the non-deliberative cause, or by creating some other one. I provided one account of why we correctly count people as being morally responsible for the actions caused by non-deliberative causes. If we think that people are morally responsible for temporally extended actions and probabilistic outcomes, by combining these we get an account of why we count as morally responsible for actions caused by our non-deliberative causes. The main upshot of this proposal is because the temporally extended actions occur alongside the creation of the non-deliberative cause it makes sense to say that someone could have interceded on these temporally extended actions. This means that people are morally responsible for the consequences of their non-deliberative causes and so we no longer have any reason according to the second objection to think that they are unfree.

The third objection I addressed was that there seems to be something wrong with indirect compatibilism if it counts directly free actions, and indirectly free actions, both paradigmatic and non-paradigmatic, as all being fully free. That's because there seems to be something very different between cases involving FCMs, P-CAPs, and NP-CAPs. In reply I

described a number of proposals that are consistent with indirect compatibilism and can accommodate the intuitions that motivate the third objection. One way we might accommodate these intuitions is with vagueness. Another way we might accommodate these intuitions is with an account of degrees of freedom that falls naturally out of indirect compatibilism. However, in order to accommodate the full range of intuitions motivating the third objection we will more than likely need to draw upon both degrees of freedom and vagueness. Irrespective of the way we choose to do it, what is important for my purposes is the fact that we can accommodate these intuitions in a manner that is compatible with indirect compatibilism.

Finally, I addressed a fourth objection according to which I have not presented any empirical evidence that might inform us of whether or not indirect compatibilism is the correct account of free action. The only empirical evidence that directly bears on a sufficient condition as to whether or not indirectly free actions are free *simpliciter* will be from experimental philosophy, which can tell us whether or not people count cases of indirect freedom as free, or upon reflection could come to judge them as free, and this only if we hold that such experimental results can be sufficient rather than an adjunct to philosophical analysis. Here I highlight one key result from my own investigations into the brain sciences: late stage negativity over the left frontopolar cortex that occurs after reward feedback under conditions of uncertainty when responding with a dominant hand. If I am right, then this is evidence that people engage FCMs even on very simple tasks provided those tasks are incentivised and performance matters. This, in turn, suggests that people do in fact update their action procedures. Now that I have finished working through the objections my defence of indirect compatibilism in this thesis is finished.

### 5.7 Conclusion: Indirect compatibilism

In this thesis I have defended a brand new account of free action: indirect (conscious control) compatibilism. This new compatibilist account of free action is the combination of two theses. The first is that the best understanding of the conceptual relationship between determinism and free will is that it is a conditional concept. The second is indirection: actions are free either when they caused by freedom-conferring mechanism, or else by sub-personal level processes influenced in various ways by freedom-conferring mechanisms (i.e., controlled action procedures). Freedom-conferring mechanisms are those conscious psychological processes that cause free actions. As such, indirect compatibilism is a heavy modification of an account of free action that I think is very obvious, yet to the best of my knowledge had not been explicitly defended in the literature on free will: conscious control.

In Chapter 2, I established that our folk concept of free will is a compatibilist concept, albeit of a unique kind. The folk concept of free will is a *conditional concept*. A conditional concept is a single concept that has an indexical component which fixes what is necessary for the concept to be satisfied, depending on what the actual world is like. If the actual world is indeterministic, and agents have libertarian powers, then these libertarian powers are what free will is and must be, else if the actual world is deterministic, and agents have their preferred compatibilist powers, then compatibilist powers are what free will is. I presented evidence from my own experimental philosophy investigation that shows that the folk concept of free will has a conditional structure, and thus shows the folk are free will compatibilists.

In Chapter 3 I addressed two challenges posed by the brain sciences to free actions, and by extension, to free will. The first challenge was the *broad challenge*: evidence from the brain sciences shows that our freedom-conferring mechanisms cause no actions, and so none of our actions are free. The second challenge was the *narrow challenge*: evidence from the brain sciences shows that our freedom-conferring mechanisms cause very few actions. I established that while the broad challenge is wrong, the narrow challenge is plausibly right, and no less problematic. That is because if our freedom-conferring mechanisms cause very few of our actions, then far fewer actions than we initially supposed will be free. Further, if not enough of our actions are free, then we do not have free will: the *discrepancy problem*. I argued that we can accommodate many of our free action judgments by acknowledging that most cases that we judge to be free are not free because they are caused directly by freedom-conferring mechanisms, but because they are caused indirectly via action procedures which those freedom-conferring mechanisms influenced in some way.

In Chapter 4 I described a number of different ways people might influence action procedures. The most obvious way people can influence action procedures is by creating them. But that is not the only route to influencing action procedures. People might be able to destroy the action procedure that they possess, or, they might be able to trigger on an action procedure they already possess, or, they might be able to monitor and turn off an action procedure, if they so choose, or, they might be able to update an action procedure they already possess to perform a different function. In each case, the resulting action procedure causes free actions because it was influenced by the agent's freedom-conferring mechanisms. I argued that indirect compatibilism is a better account of our concept of free action than is straightforward compatibilism because it is better aligned with our pre-theoretic free action judgments (our judgments about which actions are free and unfree) and vindicates the thought that we have free will.

Finally, in Chapter 5 I worked through what I think are some of the most interesting objections to indirect compatibilism. I also highlighted one finding from my own cognitive neuroscience investigations that I think is evidence of people updating an action procedure, and thus shows that an important empirical *necessary* condition is satisfied by my view. If I am right about the result, then people act freely. Further, if I am right about indirect compatibilism, then people have free will.

## References


Abramowitz, J. S., Taylor, S., and McKay, D. (2009). Obsessive-compulsive disorder. *The Lancet*. 374(9688): 419-499

Anderson, M. C., and Green, C. (2001). Suppressing unwanted memories by executive control. *Nature*. *410*(6826): 366–369.

Anderson, M. C., and Huddleston, E. (2012). Towards a cognitive and neurobiological model of motivated forgetting. In R. F. Belli (Ed.), *True and false recovered memories: Toward a reconciliation of the debate*. New York: Springer.

Angus, D. J., Latham, A. J., Harmon-Jones, E., Deliano, M., Balleine, B., and Braddon-Mitchell, D. (2017). Electrocortical components of anticipation and consumption in a monetary incentive delay task. *Psychophysiology*. 54(11): 1686-1705.

Arneson, R. J. (2003). Defending the Purely Instrumental Account of Democratic Legitimacy. *The Journal of Political Philosophy*. 11(1): 122-132.

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.

Baars, B. J. (1988) *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Badre, D., and Wagner, A. D. (2004). Selection, integration, and conflict monitoring; assessing the nature and generality of prefrontal cognitive control mechanisms. *Neuron*. 41(3): 473-487.

Banks, W. P., and Isham, E. A. (2009). We infer rather than perceive the moment we decided to act. *Psychological Science*. 20(1): 17-21.

Beaty, R. E., Benedek, M., Kaufman, S. B., and Silvia, P. J. (2015). Default and Executive Network Coupling Supports Creative Idea Production. *Scientific Reports*. 5: 10964.

Bechara, A., and Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior*. 52(2): 336-372.

Björnsson, G. (2014). Incompatibilism and "Bypassed" Agency. In A. R. Mele (ed.), *Surrounding Free Will*. New York: Oxford University Press.

Bode, S., He, A. H., Soon, C. S., Trampel, R., Turner, R., and Haynes, J. D. (2011). Tracking the unconscious generation of free decisions using ultra-high field fMRI. *PLoS One*. 6(6): e21612.

Bode, S., Murawski, C., Soon, C. S., Bode, P., Stahl, J., and Smith, P. L. (2014). Demystifying "free will": the role of contextual information and evidence accumulation for predictive brain activity. *Neuroscience Biobehavioral Review*. 47: 636-645.

Bode, S., Sewell, D. K., Lilburn, S., Forte, J. D., Smith, P. L., and Stahl, J. (2012). Predicting perceptual decision biases from early brain activity. *Journal of Neuroscience*. 32(36): 12488-12498.

Boorman, E. D., Behrens, T. E. J., Woolrich, M. W., and Rushworth, M. F. S. (2009). How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence in Favor of Alternative Courses of Action. *Neuron*. 62(5): 733-743.

Boorse, C. (1977). Health as a Theoretical Concept. *Philosophy of Science*. 44: 542–573.

Botvinick, M. M., Cohen, J. D., and Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Science*. 8(12): 539-546.

Boyd, R., Gintis, H., and Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*. 328(5978): 617-620.

Braddon-Mitchell, D. (2003). Qualia and analytical conditionals. *The Journal of Philosophy*. 100(3): 111-135.

Braddon-Mitchell, D. (2009). Naturalistic analysis and the a priori. In D. Braddon-Mitchell, and R. Nola (eds.), *Conceptual Analysis and Philosophical Naturalism*. MIT Press.

Braddon-Mitchell, D. (manuscript). Freedom and binding consequentialism.

Brass, M., and Haggard, P. (2008). The what, when, whether model of intentional action. *Neuroscientist*. 14: 319–325.

Bratman, M. (1997). Responsibility and Planning. *The Journal of Philosophy*. 1(1): 27–43.

Bratman, M. (2007). *Structures of Agency*. New York: Oxford.

Braver, T. S., and Barch, D. M. (2006). Extracting core components of cognitive control. *Trends in Cognitive Science*. 10(12): 529-532.

Bulevich, J. B., Roediger, H. L., Balota, D. A., and Butler, A. C. (2006). Failures to find suppression of episodic memories in the think/no-think paradigm. *Memory and cognition*. 34(8): 1569–1577.

Bunge, S. A., Helskog, E. H., and Wendelken, C. (2009). Left, but not right, rostrolateral prefrontal cortex meets a stringent test of the relational integration hypothesis. *NeuroImage*. 46(1): 338-342.

Bunge, S. A., Kahn, I., Wallis, J. D., Miller, E. K., and Wagner, A. D. (2003). Neural circuits subserving the retrieval and maintenance of abstract rules. *Journal of Neurophysiology*. 90: 3419–3428.

Burgess, P. W., Gonen-Yaacovi, G., and Volle, E. (2011). Functional neuroimaging studies of prospective memory: What have we learnt so far? *Neuropsychologia*. 49(8): 2246-2257.

Cameron, L. (1999). Raising the stakes in the Ultimatum Game: Experimental evidence from Indonesia. *Economic Inquiry*. 37(1): 47–59.

Carruthers, P. (2005). *Consciousness: essays from a higher-order perspective*. Oxford: Oxford University Press.

Cavanna, A. E., and Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*. 129: 564-583.

Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chalmers, D. (2004). Epistemic Two-Dimensional Semantics. *Philosophical Studies*. 118: 153–226.

Clarke, R. (1993). Towards a Credible Agent-Causal Account of Free Will. *Noûs*. 27(2): 191–203.

Clarke, R. (1996). Agent Causation and Event Causation in the Production of Free Action. *Philosophical Topics*. 24(2): 19-48.

Cocchi, L., Zalesky, A., Fornito, A., and Mattingley, J. B. (2013). Dynamic cooperation and competition between brain systems during cognitive control. *Trends in Cognitive Science*. 17(10): 493-501.

Cole, M. W., Reynolds, J. R., Power, J. D., Repovs, G., Anticevic, A., and Brave, T. S., (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature Neuroscience*. 16(9): 1348-1355.

Connolly, J. D., Andersen, R. A., and Goodale, M. A. (2003). FMRI evidence for a 'parietal reach region' in the human brain. *Experimental Brain Research*. 153(2): 140-145.

Corberos, P. (2011). Paraconsistent Vagueness: A Positive Argument. *Synthese* 183(2): 211–227.

Culham, J. C., Brandt, S. A., Cavanagh, P., Kanwisher, N. G., Dale, A. M., and Tootell, R. B. H. (1998). Cortical fMRI activation produced by attentive tracking of moving targets. *Journal of Neurophysiology*. 80: 2657–2670.

Cunnington, R., Windischberger, C., Deecke, L., and Moser, E. (2002). The preparation and execution of self-initiated and externally-triggered movement: a study of event-related fMRI. *NeuroImage*. 15:373-385.

Cunnington, R., Windischberger, C., Deecke, L., and Moser, E. (2003). The preparation and readiness for voluntary movement: a high-field event-related fMRI study of the Bereitschafts-BOLD response. *NeuroImage*. 20:404–412.

Davidson, D. (1980). *Essays on Actions and Events*. Oxford: Oxford University Press.

de Quervain, D. J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., and Fehr, E. (2004). The neural basis of altruistic punishment. *Science*. 305(5688): 1254-1258.

Deecke, L., Grözinger, B., and Kornhuber, H. H. (1976). Voluntary finger movement in man: cerebral potentials and theory. *Biological Cybernetics*. 23:99-119.

Dehaene, S., and Naccache, L. (2001). Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework. *Cognition* 79(1): 1–37.

Dehaene, S., Changeux, J., Naccache, L., Sackur, J., and Sergent, Cl. (2006). Conscious, Preconscious, and Subliminal Processing: A Testable Taxonomy. *Trends in Cognitive Sciences*. 10(5): 204–211.

Dehaene, S., Kerszberg, M., and Changeux, J. (1998). A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks. *Proceedings of the National Academy of Sciences of the United States of America*. 95(24): 14529–14534.

Dehaene, S., Piazza, M., Pinel, P., and Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology*. 20(3): 487-506

Dennett, D. C. (1968). *Content and consciousness*. London: Routledge and Kegan Paul.

Dennett, D. C. (1978). *Brainstorms*. Montgomery, VT: Bradford Books.

Desmurget, M., Reilly, K. T., Richard, N., Szathmari, A., Mottolese, C., and Sirigu, A. (2009). Movement intention after parietal cortex stimulation in humans. *Science*. 324(5928): 811-813.

Dosenbach, N. U., et al. (2007). Distinct brain networks for adaptive and stable task control in humans. *Proceedings of the National Academy of Sciences of the United States of America*. 104(26): 11073-11078.

Dunaway, B., Edmonds, A., and Manley, D. (2013). The Folk Probably do Think What you Think They Think. *Australasian Journal of Philosophy*. 91(3): 421-441.

Dunn, M. J., and Hardegree, G. M. (2001). *Algebraic Methods in Philosophical Logic*. Oxford: Science Publications.

Eimer, M. (1998). The lateralized readiness potential as an on-line measure of selective response activation. *Behavior Research Methods, Instruments, & Computers*. 30:146-156.

Ekstrom, L. (2000). *Free Will: A Philosophical Study*. Boulder, CO: Westview Press.

Ekstrom, L. (2002). Libertarianism and Frankfurt-style cases. In R. Kane (ed.), *The Oxford Handbook of Free Will, 2nd Ed.* New York: Oxford University Press.

Ekstrom, L. (2003). Free Will, Chance, and Mystery. *Philosophical Studies*. 113(2): 153-180.

Erdelyi, M. H. (2006). The unified theory of repression. *Behavioral and Brain Sciences*. 29: 499–551.

Ericksen, K. P., and Horton, H. (1992). "Blood feuds": Cross-cultural variations in kin group vengeance. *Behavior Science Research*. 26(1-4): 57-85.

Falk R., and Konold C. (1997). Making sense of randomness: implicit encoding as a bias for judgment. *Psychological Review*. 104: 301–318.

Fara, D. G. (2008). Profiling Interest-Relativity. *Analysis*. 68(4): 326–335.

Farrer, C., and Frith, C. D. (2002). Experiencing oneself vs. another person as being the cause of an action: the neural correlates of the experience of agency. *NeuroImage*. 15: 596–603.

Farrow, T. F. D., et al. (2001). Investigating the functional anatomy of empathy and forgiveness. *Neuroreport*. 12: 2433–2438.

Fehl, K., Sommerfeld, R. D., Semmann, D., Krambeck, H., and Milinski, M. (2012). I Dare You to Punish Me - Vendettas in Games of Cooperation. *PLoS ONE*. 7(9): e45093.

Fehr, E., and Fischbacher, U. (2003). The nature of human altruism. *Nature*. 425: 785-791.

Fehr, E., and Fischbacher, U. (2005). The Economics of Strong Reciprocity. In H. Gintis, R. Boyd, S. Bowles, and E. Fehr (eds.), *Moral Sentiments and Material Interests*. Cambridge, Mass.: MIT Press.

Fellows, L. K. (2007). The Role of Orbitofrontal Cortex in Decision Making. *Annals of The New York Academy Of Sciences*. 1121(1): 421-430.

Fine, K. (1975). Vagueness, Truth and Logic. *Synthese*. 30(3–4): 265–300.

Fischer, J. M. (2004). Responsibility and Manipulation. *Journal of Ethics*. 8(2): 145–77.

Fischer, J. M., and Ravizza, M. (1998). *Responsibility and Control: An Essay on Moral Responsibility*. Cambridge: Cambridge University Press.

Fischer, J. M., and Tognazzini, N. A. (2009). The Truth about Tracing. *Noûs*. 43(3): 531-556.

Fischer, J. M., and Tognazzini, N. A. (2012). The Triumph of Tracing. In J. M. Fischer (ed.), *Deep Control*. Oxford University Press.

Fischer, S., Diekelmann, S., and Born, J. (2011). Sleep's role in the processing of unwanted memories. *Journal of sleep research*. *20*(2): 267–274.

Fletcher, P. C., Frith, C. D., Baker, S. C., Shallice, T., Frackowiak, R. S, and Dolan, R. J. (1995). The mind's eye—precuneus activation in memory-related imagery. *NeuroImage*. 2:195–200.

Frank, M. J., Woroch, B. S., and Curran, T. (2005). Error-Related Negativity Predicts Reinforcement Learning and Conflict Biases. *Neuron*. 47(4): 495-501.

Frankfurt, H. G. (1969). Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy*. 66(23): 829-839.

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*. 68: 5–20.

Frankfurt, H. G. (1987). Identification and Wholeheartedness. In F. Schoeman (ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. Cambridge: Cambridge University Press.

Frankfurt, H. G. (1994). Autonomy, Necessity, and Love. In H.F., Fulda, and R.P. Horstman (eds.) *Vernunftbegriffe in der Moderne: Stuttgart Hegel-Kongress 1993*. Stuttgart: Klett-Cotta.

Fried, I., Mukamel, R., and Kreiman, G. (2011). Internally generated preactivation of single neurons in human medial frontal cortex predicts volition. *Neuron*. 69(3): 548-562.

Gächter, S., and Herrmann, B. (2008). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B*. 364(1518): 791-806.

Gardner, A., and West, S. A. (2014). Inclusive fitness: 50 years on. *Philosophical Transactions of the Royal Society B*. 36: 20130356.

Gazzaniga, M. S. (2011). *Who's in Charge? Free Will* and the *Science of the Brain*. New York: Ecco.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., and Donchin, E. (1993). A Neural System for Error Detection and Compensation. *Psychological Science*. 4(6): 385-390.

Ghaem, O., et al. (1997). Mental navigation along memorized routes activates the hippocampus, precuneus, and insula. *Neuroreport*. 8: 739–744.

Gilboa, A., Winocur, G., Grady, C. L., Hevenor, S. J., and Moscovicth, M. (2004). Remembering our past: functional neuroanatomy of recollection of recent and very remote personal events. *Cerebral Cortex*. 14: 1214–1225.

Ginet, C. (1990). *On Action*. Cambridge: Cambridge University Press.

Gintis, H., Smith, E. A., and Bowles, S. (2001). Costly Signaling and Cooperation. *Journal of Theoretical Biology*. 213: 103-119.

Godfrey-Smith, P. (1994). A Modern History Theory of Functions. *Noûs*, 28: 344–362.

Goldman, A. (1970). *A Theory of Human Action*. Englewood Cliffs, NJ: Prentice-Hall.

Gottwald, S. (2001). *A Treatise on Many-valued Logic*. Baldock: Research Studies Press.

Gouldner, A. W. (1960). The Norm of Reciprocity: A Preliminary Statement. *American Sociological Review*. 25(2): 161-178.

Grabner, R. H., Ansari, D., Koschutnig, K., Reishofer, G., Ebner, F., and Neuper, C. (2009). To retrieve or to calculate? Left angular gyrus mediates the retrieval of arithmetic facts during problem solving. *Neuropsychologia*. 47(2):604–8

Graff, D. (2000). Shifting Sands: An Interest-Relative Theory of Vagueness. *Philosophical Topics*. 28(1): 45–81.

Greene, J., and Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society B*. 359: 1775-1785.

Greicius, M. D., Krasnow, B., Reiss, A. L., and Menon, V. (2003). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*. 100(1): 253-258.

Griffiths, P. E. (1993). Functional Analysis and Proper Function. *British Journal for Philosophy of Science*. 44: 409–422.

Griffiths, P. E., and Matthewson, J. (2018). Evolution, Dysfunction, and Disease: A Reappraisal. *The British Journal for the Philosophy of Science*. 69(2): 301-327.

Guala, F. (2012). Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*. 35(1): 1-15.

Gurven, M., Zanolini, A., and Schniter, E. (2008). Culture sometimes matters: intra-cultural variation in pro-social behavior among Tsimane Amerindians. *Journal of Economic Behavior and Organization*. 67(3-4): 587-607.

Güth, W., Schmittberger, R., and Schwarz, B. (1982). An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior and Organization*. 3: 367-388.

Haggard, P. (2011). Neuroethics of free will. In J. Illes, and B. J. Sahakian (eds.), *Oxford Handbook of Neuroethics*. New York: Oxford University Press.

Haggard, P., Clark, S., and Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*. 5(4): 382-385.

Haggard, P., and Eimer, M. (1999). On the relation between brain potentials and the awareness of voluntary movements. *Experimental Brain Research*. 126: 128-133.

Haji, I. (2000). Libertarianism and the Luck Objection. *Journal of Ethics*. 4: 329–37.

Hallett, M. (2007). Volitional control of movement: the physiology of free will. *Clinical Neurophysiology*. 118(6): 1179-1192.

Hamilton, W. D. (1964). The Genetical Evolution of Social Behaviour. I. *Journal of Theoretical Biology*. 7: 1-16.

Hart, H. L. A. (1968). *Punishment and Responsibility*. Oxford University Press.

Hawthorne, J. (2002). Advice for Physicalists. *Philosophical Studies*. 109: 17-52.

Hayne, H., Garry, M., and Loftus, E. F. (2006). On the continuing lack of scientific evidence for repression. *Behavioral and Brain Sciences*. 29(5): 521-522.

Henrich, J., Boyd, R., Bowls, S., Camerer, C., Fehr, E., and Gintis, H. (eds.). (2004). *Foundations of human sociality*. New York: Oxford University Press.

Hobart, R. E. (1934). Free Will as Involving Determinism and Inconceivable Without It". *Mind*. 43: 1–27.

Hobbes, T. (1651/1997). *Leviathan*, R.E. Flatman and D. Johnston (eds.), New York: W.W. Norton & Co.

Holroyd, C. B., and Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*. 109(4): 679-709.

Holroyd, C. B., Coles, M. G. H., and Nieuwenhuis, S. (2002). Medial prefrontal cortex and error potentials. *Science*. 296(5573): 1610-1611.

Hume, D. (1740/1978). *A Treatise of Human Nature*, P.H. Nidditch (ed.), Oxford: Clarendon Press.

Hyde, D. (2008). *Vagueness, Logic and Ontology*. Aldershot: Ashgate.

Hyde, D., and Colyvan, M. (2008). Paraconsistent Vagueness: Why Not? *Australasian Journal of Logic*. 6: 107–121.

Jackson, F. (1998). *From Metaphysics to Ethics*. Oxford: Clarendon Press.

Jahanshahi, M. (1998). Willed action and its impairments. *Cognitive Neuropsychology*. 15: 483–533.

Johnson, S. H. (2000). Thinking ahead: The case for motor imagery in prospective judgements of prehension. *Cognition*. 74:33–70.

Kamm, F. M. (2007). *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. New York: Oxford University Press.

Kane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.

Kane, R. (2005). *A contemporary introduction to free will*. New York: Oxford University Press.

Karl, J. M., and Whishaw, I. Q. (2013). Different Evolutionary Origins for the Reach and the Grasp: An Explanation for Dual Visuomotor Channels in Primate Parietofrontal Cortex. *Frontiers in Neurology*. 4: 208.

Keas, M. N. (2018). Systematizing the theoretical virtues. *Synthese*. 195(6): 2761-2793.

Keefe, R. (2000). *Theories of Vagueness*, Cambridge: Cambridge University Press.

Kihlstrom, J. (2002). No need for repression. *Trends in Cognitive Sciences*. 6: 502.

Kircher, T. T. J., Brammer, M., Bullmore, E., Simmons, A., Bartels, M., and David, A. S. (2002). The neural correlates of intentional and incidental self-processing. *Neuropsychologia*. 40: 683–692.

Kircher, T. T. J., et al. (2000). Towards a functional neuroanatomy of self-processing: effects of faces and words. *Cognitive Brain Research*. 10: 133–144.

Kjaer, T. W., Nowak, M., Kjaer, K. W., Lou, A. R., and Lou, H. C. (2001). Precuneus-prefrontal activity during awareness of visual verbal stimuli. *Conscious and Cognition*. 10: 356–365.

Knauff, M., Fangmeier, T., Ruff, C. C., and Johnson-Laird, P.N. (2003). Reasoning, models, and images: behavioral measures and cortical activity. *Journal of Cognitive Neuroscience*. 15: 559–573.

Koechlin, E., Corrado, G., Pietrini, P., and Grafman, J. (2000). Dissociating the role of the medial and lateral anterior prefrontal cortex in human planning. *Proceedings of the National Academy of Sciences of the United States of America*. 97(13): 7651-7656.

Kriegel, U., 2003. Consciousness as intransitive self-consciousness: two views and an argument. *Canadian Journal of Philosophy*. 33: 103–132.

Lafargue, G., and Duffau, H. (2008). Awareness of intending to act following parietal cortex resection. *Neuropsychologia*. 46(11): 2662-2667.

Lages, M., Boyle, S. C., and Jaworska, K. (2013). Flipping a coin in your head without monitoring outcomes? Comments on predicting free choices and a demo program. *Frontiers in Psychology*. 4: 925.

Lamme, V. A. F. (2006). Towards a True Neural Stance on Consciousness. *Trends in Cognitive Sciences*. 10(11): 494–501.

Lamme, V. A. F. (2010). How Neuroscience Will Change Our View on Consciousness. *Cognitive Neuroscience*. 1(3): 204–220.

Latham, A. J., Ellis, C., Chan, L., & Braddon-Mitchell, D. (2017). The Validation of Consciousness Meters: The Idiosyncratic and Intransitive Sequence of Conscious Levels. *Journal of Consciousness Studies*. 24(3-4): 103-111.

Lau, H. C., Rogers, R. D., and Passingham, R. E. (2007). Manipulating the Experienced Onset of Intention after Action Execution. *Journal of Cognitive Neuroscience*. 19: 81–90.

Leech, R., and Sharp, D. J. (2014). The role of the posterior cingulate cortex in cognition and disease. *Brain*. 137(1): 12-32.

Leech, R., Kamourieh, S., Beckmann, C. F., and Sharp, D. J. (2011). Fractionating the Default Mode Network: Distinct Contributions of the Ventral and Dorsal Posterior Cingulate Cortex to Cognitive Control. *The Journal of Neuroscience*. 31(9): 3217-3224.

Levy, N. (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford: Oxford University Press.

Lewis, D. (1986). *On the Plurality of Worlds*. Oxford: Basil Blackwell.

Lewis, D. (1995). Should a materialist believe in qualia? *Australasian Journal of Philosophy*. 73(1): 140-144.

Liao, S. (2016). Are philosophers good intuition predictors? *Philosophical Psychology*. 29(7): 1004-1014.

Libet, B. (1985). Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*. 8: 529-566.

Libet, B. (1999). Do we have free will? *Journal of Consciousness Studies*. 6(8-9): 47-57.

Libet, B. (2004). *Mind Time: The Temporal Factor in Consciousness*. Cambridge, MA: Harvard University Press.

Libet, B., Gleason, C. A., Wright, E. W., and Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. *Brain*. 106(3): 623-42.

Lopes, L. (1982). Doing the impossible: a note on induction and experience of randomness. *Journal of Experimental Psychology*. 8: 626–636.

Lotem, A., Fishman, M. A., and Stone, L. (2003). From reciprocity to unconditional altruism through signalling benefits. *Proceedings of the Royal Society B*. 270(1511): 199-205.

Malouin, F., Richards, C. L., Jackson, P. L., Dumas, F., and Doyon, J. (2003). Brain activations during motor imagery of locomotor-related tasks: a PET study. *Human Brain Mapping*. 19: 47–62.

Margulies, D. S., et al. (2009). Precuneus shares intrinsic functional architecture in humans and monkeys. *Proceedings of the National Academy of Sciences of the United States of America*. 106(47): 20069-20074.

Markosian, N. (1999). A Compatibilist Version of the Theory of Agent Causation. *Pacific Philosophical Quarterly*. 80: 257–77.

Markosian, N. (2012). Agent Causation as the Solution to all the Compatibilist's Problems. *Philosophical Studies*. 157: 383–98.

Matthew, S., and Boyd, R. (2014). The cost of cowardice: punitive sentiments towards free riders in Turkana raids. *Evolution and Human Behavior*. 35(1): 58-64.

Matthewson, J., and Griffiths, P. E. (2017). Biological Criteria of Disease: Four Ways of Going Wrong. *Journal of Medical Philosophy*. 42(4): 447-466.

McKenna, M. (2005). The relationship between autonomous and morally responsible agency. In J. Stacey Taylor (ed.), *Personal Autonomy: New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy*. Cambridge University Press.

McKenna, M. (2008). Putting the Lie on the Control Condition for Moral Responsibility. *Philosophical Studies*. 139(1): 29–37.

McKenna, M. and Coates, J. D. (2015). Compatibilism. In E. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu

Mele, A. R. (2006). *Free Will and Luck*. Oxford: Oxford University Press

Mele, A. R. (1995). *Autonomous Agents*. New York: Oxford University Press.

Mele, A. R. (1996). Soft Libertarianism and Frankfurt-Style Scenarios. *Philosophical Topics*. 24(2): 123-141.

Mele, A. R. (1999). Kane, Luck, and the Significance of Free Will. *Philosophical Explorations*. 2: 96–104.

Mele, A. R. (2010) Scientific Skepticism about Free Will. In T. Nadelhoffer, E. Nahmias, and S. Nichols (eds.), *Moral Psychology: Classical and Contemporary Readings*. Blackwell.

Mele, A. R. (2014). *Free: Why Science Hasn't Disproved Free Will*. New York: Oxford University Press.

Miller, J., Shepherdson, P., and Trevena, J. (2011). Effects of clock monitoring on electroencephalographic activity: is unconscious movement initiation an artifact of the clock? *Psychological Science*. 22(1): 103-109.

Montague, P. R. (2008). Free will. *Current Biology*. 18(14): 584-585.

Murray, D., and Nahmias, E. (2014). Explaining Away Incompatibilist Intuitions. *Philosophy and Phenomenological Research*. 88(2): 434-467.

Mylopoulos, M. I., and Lau, H. (2014). Naturalizing Free Will. In A. R. Mele (ed.), *Surrounding Free Will: Philosophy, Psychology, Neuroscience*. Oxford University Press.

Nagahama, Y., et al. (1999). Transient neural activity in the medial superior frontal gyrus and precuneus time locked with attention shift between object features. *NeuroImage*. 10: 193–199.

Nahmias, E. (2002). When Consciousness Matters: A Critical Review of Daniel Wegner's The Illusion of Conscious Will. *Philosophical Psychology*. 15(4): 527-542.

Nahmias, E. (2011). Intuitions about Free Will, Determinism, and Bypassing. In R. Kane (ed.), *The Oxford Handbook of Free Will, 2nd Ed.* New York: Oxford University Press.

Nahmias, E. and Murray, D. (2010). Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions. In J. Aguilar, A. Buckareff, and K. Frankish (eds.), *New Waves in Philosophy of Action*. Palgrave-Macmillan.

Nahmias, E., Coates, J. D., and Kvaran, T. (2007). Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions. *Midwest Studies in Philosophy*. 31(1): 214-242.

Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (2005). Surveying freedom: folk intuitions about free will and moral responsibility. *Philosophical Psychology*. 18: 561-584.

Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (2006). Is incompatibilism intuitive? *Philosophy and Phenomenological Research*. 73(1): 28-53.

Nichols, S. B. (2004) The Folk Psychology of Free Will: Fits and Starts. *Mind & Language*. 19(5): 473-502.

Nichols, S. B., and Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs*. 41(4): 663-685.

Nikiforakis, N., and Engelmann, D. (2011). Altruistic punishment and the threat of feuds. *Journal of Economic Behavior and Organization*. 78(3): 319-332.

Noreen, S., and MacLeod, M. D. (2013). It's all in the detail: Intentional forgetting of autobiographical memories using the autobiographical think/no-think task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. *39*(2): 375–393.

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*. 314(5805): 1560-1563.

Nowell-Smith, P. (1948). Free Will and Moral Responsibility. *Mind*. 55: 45–61

O'Connor, T. and Franklin, C. (2018). Free Will. In E. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu

O'Connor, T. (1995). *Agents, Causes, and Events: Essays on Indeterminism and Free Will*. New York: Oxford University Press.

O'Connor, T. (2000). *Persons and Causes: the Metaphysics of Free Will*, New York: Oxford University Press.

O'Connor, T. (2005). Freedom With a Human Face. *Midwest Studies in Philosophy*. 29(1): 207-227.

O'Connor, T. (2009). Degrees of freedom. *Philosophical Explorations*. 12(2): 119-125.

Ochsner, K. N., et al. (2004). Reflecting upon feelings: an MRI study of neural systems supporting the attribution of emotion to self and other. *Journal of Cognitive Neuroscience*. 16: 1746–1772.

Ogiso, T., Kobayashi, K., and Sugishita, M. (2000). The precuneus in motor imagery: a magnetoencephalographic study. *Neuroreport*. 11: 1345–1349.

Okuda, J., et al. (2003). Thinking of the future and past: the roles of the frontal pole and the medial temporal lobes. *NeuroImage*. 19(4): 1369-1380.

Oliveira, F. T. P., Diedrichsen, J., Verstynen, T., Duque, J., and Ivry, R. B. (2010). Transcranial magnetic stimulation of posterior parietal cortex affects decisions of hand choice. *Proceedings of the National Academy of Sciences of the United States of America*. 107(41): 17751-17756.

Orr, J. M., and Carrasco, M. (2011). The Role of the Error Positivity in the Conscious Perception of Errors. *The Journal of Neuroscience*. 31(16): 5891-5892.

Otsuka, M. (1998). Incompatibilism and the avoidability of blame. *Ethics*. 108(4): 685-701.

Pereboom, D. (2001). *Living Without Free Will*, Cambridge: Cambridge University Press.

Pink, T. (2004). *Free will: A very short introduction*. Oxford: Oxford University Press.

Pockett, S. P. (2007). The concept of free will: philosophy, neuroscience and the law. *Behavioral Sciences & the Law*. 25(2): 281-293.

Praamstra, P., Stegeman, D. F., Horstink, M. W., and Cools, A. R. (1996). Dipole source analysis suggests selective modulation of the supplementary motor area contribution to the readiness potential. *Electroencephalography and Clinical Neurophysiology*. 98: 468-477.

Quackenbush, S. L. (2010). General Deterrence and International Conflict: Testing Perfect Deterrence Theory. *International Interactions*. 36: 60-85.

Rachels, J. (1975). Active and passive euthanasia. *New England Journal of Medicine*. 292(2): 78-80.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America*. 98(2): 676-682.

Railton, P. (2014). The affective dog and its rational tale: intuition and attunement. *Ethics*. 124(4): 813-859.

Ramnani, N., and Owen, A. M. (2004). Anterior prefrontal cortex: insights into function from anatomy and neuroimaging. *Nature Reviews Neuroscience*. 5(3): 184-194.

Rauchhaus, R. (2009). Evaluating the Nuclear Peace Hypothesis: A Quantitative Approach. *The Journal of Conflict Resolution*. 53(2): 258-277.

Rawls, J. (1971) *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rescher, N. (2009). *Unknowability: An Enquiry into the Limits of Knowledge*, New York: Lexington Books.

Reynolds, J. R., McDermott, K. B., and Braver, T. S. (2006). A Direct Comparison of Anterior Prefrontal Cortex Involvement in Episodic Retrieval and Integration. *Cerebral Cortex*. 16(4): 519-528.

Rose, D., and Nichols, S. (2013). The Lesson of Bypassing. *Review of Philosophy and Psychology*. 4(4): 599-619.

Rosenthal, D. (2005). *Consciousness and Mind*. Oxford: Oxford University Press.

Roskies, A. L. (2010). How does neuroscience affect our conception of volition? *Annual Review of Neuroscience*. 33: 109-130.

Roskies, A. L. and Nichols, S. B. (2008). Bring moral responsibility down to earth. *The Journal of Philosophy*. 105(7): 371-388.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., and Zamir, S. (1991). Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh and Tokyo: An Experimental Study. *The American Economic Review*. 81(5): 1068-1095.

Rushworth, M. F., Noonan, M. P., Boorman, E. D., Walton, M. E., and Behrens, T. E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron*. 70(6): 1054-1069.

Ryckman, N. A., Addis, D. R., Latham, A. J., and Lambert, A. J. (2018). Forget about the future: effects of thought suppression on memory for imaginary emotional episodes. *Cognition and Emotion*. 32(1): 200-206.

Ryckman, N. A., Addis, D. R., Latham, A. J., and Lambert, A. J. (manuscript). Suppressing Imagined Events: Effects of neuroticism and emotional valence on memory for imagined autobiographical episodes.

Sakai, K., and Passingham, R. E. (2003). Prefrontal interactions reflect future task operations. *Nature Neuroscience*. 6: 75–81.

Sakai, K., and Passingham, R. E. (2006). Prefrontal set activity predicts rule-specific neural processing during subsequent cognitive performance. *Journal of Neuroscience*. 26(4): 1211-1218.

Satoh, M., Takeda, K., Nagata, K., Hatazawa, J., and Kuzuhara, S. (2001). Activated brain regions in musicians during an ensamble: a PET study. *Cognitive Brain Research*. 12: 101–108.

Schlegel, A., Alexander, P., Sinnott-Armstrong, W., Roskies, A., Tse, P. U., and Wheatley, T. (2013). Barking up the wrong free: readiness potentials reflect processes independent of conscious will. *Experimental Brain Research*. 229(3): 329-335.

Schlick, M. (1935). Facts and Propositions. *Analysis*. 2(5): 65-70.

Schultze-Kraft, M., Birman, D., Rusconi, M., Allefeld, C., Görgen, K., Dähne, S., Blankertz, B., and Haynes, J. D. (2016). The point of no return in vetoing self-initiated movements. *Proceedings of the National Academy of Sciences of the United States of America*. 113(4): 1080-1085.

Schurger, A., Sitt, J. D., and Dehaene, S. (2012). An accumulator model for spontaneous neural activity prior to self-initiated movement. *Proceedings of the National Academy of Sciences of the United States of America*. 109(42): 2904-2913.

Sechser, T. S., and Fuhrmann, M. (2013). Crisis Bargaining and Nuclear Blackmail. *International Organization*. 67(1): 173-195.

Shepherd, J. (2012). Free Will and Consciousness: Experimental Studies. *Consciousness and Cognition*. 21(2): 915–927.

Shepherd, J. (2015). Consciousness, free will, and moral responsibility: Taking the folk seriously. *Philosophical Psychology*. 28(7): 929-946.

Shibasaki, H., and Hallett, M. (2006). What is the Bereitschaftspotential? *Clinical Neurophysiology*. 117: 2341-2356.

Shoemaker, D. W. (2003). Caring, Identification, and Agency. *Ethics*. 114(1): 88-118.

Shramko, Y., and Wansing, H. (2011). *Truth and Falsehood: An Inquiry into Generalized Logical Values*. New York: Springer.

Simon, O., Mangin, J. F., Cohen, L., Le Bihan, D., and Dehaene, S. (2002). Topographical layout of hand, eye, calculation, and language-related areas in the human parietal lobe. *Neuron*. 33: 475–487.

Sirigu, A., Daprati, E. Ciancia, S., Giraux, P., Nighoghossian, N., Posada, A., and Haggard, P. (2004). Altered awareness of voluntary action after damage to the parietal cortex. *Nature Neuroscience*. 7(1): 80-84.

Sirotin, Y. B., and Das, A. (2009). Anticipatory haemodynamic signals in sensory cortex not predicted by local neuronal activity. *Nature*. 457: 475-479.

Smallwood, J., Brown, K., Baird, B., and Schooler, J. W. (2012). Cooperation between the default mode network and the frontal-parietal network in the production of an internal train of thought. *Brain Research*. 1428: 60-70.

Smart, J. J. C. (1961). Free-will, Praise and Blame. *Mind*. 70: 291– 306.

Smith, E. A., and Bird, R. (2000). Turtle hunting and tombstone opening: public generosity as costly signaling. *Evolution and Human Behavior*. 21:245–261.

Smith, K. (2011). Neuroscience vs philosophy: Taking aim at free will. *Nature*. 477: 23-25.

Smith, N. J. J. (2008). *Vagueness and Degrees of Truth*. New York: Oxford University Press.

Snyder, A. Z., and Raichle, M. E. (2012). A brief history of the resting state: the Washington University perspective. *NeuroImage*. 62(2): 902-910.

Soon, C. S., Brass, M., Heinze, H. J., and Haynes, J. D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*. 11(5): 543-545.

Soon, C. S., He, A. H., Bode, S., and Haynes, J. D. (2013). Predicting free choices for abstract intentions. *Proceedings of the National Academy of Sciences of the United States of America*. 110(15): 6217-6222.

Sorensen, R. A. (1988). *Blindspots*. Oxford: Clarendon Press.

Sorensen, R. A. (2001). *Vagueness and Contradiction*. New York: Oxford University Press.

Stalnaker, R. (1978). Assertion. *Syntax and Semantics*. 9: 315–332.

Stalnaker, R. (2002). What is it Like to be a Zombie? In T. Gendler, and J. Hawthorne (eds.), *Conceivability and Possibility*. New York: Oxford University Press.

Stanescu-Cosson, R., Pinel, P., van De Moortele, P. F., Le Bihan, D., Cohen, L., and Dehaene, S. (2000). Understanding dissociations in dyscalculia: a brain imaging study of the impact of number size on the cerebral networks for exact and approximate calculation. *Brain*. 123(11): 2240-2255.

Steinhauser, M., and Yeung, N. (2010). Decision Processes in Human Performance Monitoring. *The Journal of Neuroscience*. 30(46): 15643-15653.

Strawson, G. (1986) *Freedom and Belief*. Oxford: Clarendon Press.

Strawson, G. (1994). The Impossibility of Ultimate Moral Responsibility. *Philosophical Studies*. 75(1–2): 5–24.

Strawson, P. (1962). Freedom and resentment. In G. Watson (ed.), *Proceedings of the British Academy, Volume 48:1962*. Oxford: Oxford University Press.

Suchan, B., et al. (2002). Hemispheric dissociation of visuo-spatial processing and visual rotation. *Cognitive Brain Research*. 136: 533–544.

Tallis, R. (2011). *Aping mankind, neuromania, Darwinist and the misrepresentation of humanity*. Acumen: Durham.

Tononi, G. (2004). An Information Integration Theory of Consciousness. *BMC Neuroscience*, 5: 42.

Tononi, G. (2008). Consciousness as Integrated Information: A Provisional Manifesto. *The Biological Bulletin*. 215(3): 216–242.

Tooby, J., and Cosmides, L. (1992). The psychological foundations of culture. In J.H. Barkow, L. Cosmides, and J. Tooby (eds.), *The adapted mind: Evolutionary psychology and the generation of culture*. New York: Oxford University Press.

Tosoni, A., Galati, G., Romani, G. L., and Corbetta, M. (2008). Sensory-motor mechanisms in human parietal cortex underlie arbitrary visual decisions. *Nature Neuroscience*. 11(12): 1446-1453.

Treisman, M., and Faulkner, A. (1987). Generation of random sequences by human subjects: cognitive operations or psychophysical process. *Journal of Experimental Psychology: General*. 116: 337–355

Trivers, R. L. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology*. 46(1): 35-57.

Tye, M. (1994). Sorites Paradoxes and the Semantics of Vagueness. *Philosophical Perspectives*, 8: 189–206.

Ullsperger, M., Harsay, H. A., Wessel, J. R., and Ridderinkhof, K. R. (2010). Conscious perception of errors and its relation to the anterior insula. *Brain Structure and Function*. 214(5-6): 629-643.

van Harskamp, N. J., Rudge, P., and Cipolotti, L. (2002). Are multiplication facts implemented by the left supramarginal and angular gyri? *Neuropsychologia*. 40(11): 1786-1793.

van Inwagen, P. (1975). The Incompatibility of Free Will and Determinism. *Philosophical Studie*s. 27(3): 185-199.

van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Oxford University Press.

van Inwagen, P. (1993) *Metaphysics*. Boulder: Westview Press.

van Inwagen, P. (2008). How to Think about the Problem of Free Will. *The Journal of Ethics*. 12(3-4): 327-341.

Vargas, M. (2005). The Trouble With Tracing. *Midwest Studies in Philosophy*, 29: 269-291.

Vargas, M. (2013). *Building Better Beings*. Oxford University Press.

Varzi, A. C. (2001). Vagueness, Logic and Ontology. *The Dialogue. Yearbooks for Philosophical Hermeneutics*. 1: 135–154.

Vaughan, H. G., Costa., L. D., and Ritter, W. (1968). Topography of the human motor potential. *Electroencephalography and Clinical Neurophysiology*. 25: 1-10.

Vincent, J. L., Kahn, I., Snyder, A. Z., Raichle, M. E., and Buckner, R. L. (2008). Evidence for a Frontoparietal Cortical System Revealed by Intrinsic Functional Connectivity. *Journal of Neurophysiology*. 100(6): 3328-3342.

Vocat, R., Pourtois, G., and Vuilleumier, P. (2008). Unavoidable errors: A spatio-temporal analysis of time-course and neural sources of evoked potentials associated with error processing in a speeded task. *Neuropsychologia*. 46(10): 2545-2555.

Wakefield, J. C. (1992). The Concept of Mental Disorder. *American Psychologist*. 47: 373–388.

Watson, G. (1975). Free Agency. *The Journal of Philosophy*. 72: 205–20.

Watson, G. (1987). Responsibility and the Limits of Evil: Variations on a Strawsonian Theme. In F. Schoeman (ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. Cambridge: Cambridge University Press.

Weber, Z., and Colyvan, M. (2010). A Topological Sorites. *The Journal of Philosophy*. 107(6): 311–325.

Wegner, D. M. (2002). *The Illusion of Conscious Will*. MIT Press.

West, R. (2011). The temporal dynamics of prospective memory: A review of the ERP and prospective memory literature. *Neuropsychologia*. 49(8): 2233-2245.

West, R., and Moore, K. (2002). Adjustments of Cognitive Control in Younger and Older Adults. *Cortex*. 41: 570-581.

Williamson, T. (1994). *Vagueness*. London: Routledge.

Williamson, T. (2000). *Knowledge and its Limits*, Oxford: Oxford University Press.

Woollard, F. (2015). *Doing and Allowing Harm*. Oxford: Oxford University Press.

Yamagishi, T., et al. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proceedings of the National Academy of Sciences of the United States of America*. 109(5): 20364-20368.

Yoshida, W., and Ishii, S. (2006). Resolution of uncertainty in prefrontal cortex. *Neuron*. 50(5): 781-789.

Zadeh, L. A. (1965). Fuzzy Sets. *Information and Control*. 8(3): 338–353.

Zapparoli, L., Seghezzi, S., and Paulesu, E. (2017). The What, the When, and the Whether of Intentional Action in the Brain: A Meta-Analytical Review. *Frontiers in Human Neuroscience*. 11: 238.

# Appendix A

# Electrocortical Components of Anticipation and Consumption in a Monetary Incentive Delay Task

Douglas J. Angus*[1]

Andrew J. Latham*[2, 3]

Eddie Harmon-Jones[1]

Matthias Deliano[4]

Bernard Balleine[1]

David Braddon-Mitchell[2]

Manuscript accepted for publication in Psychophysiology.

*Co-first authors

[1]School of Psychology
University of New South Wales
NSW 2052, Australia

[2]School of Philosophy
University of Sydney
NSW 2006, Australia

[3]Brain & Mind Research Institute
University of Sydney
NSW 2050, Australia

[4]Department Systemphysiology
Leibniz Institute for Neurobiology
Brenneckestraße 6
39118 Magdeburg, Germany

**Abstract**

In order to improve our understanding of the components that reflect functionally important processes during reward anticipation and consumption, we used principle components analyses (PCA) to separate and quantify averaged ERP data obtained from each stage of a modified monetary incentive delay (MID) task. Although a small number of recent ERP studies have reported that reward and loss cues potentiate ERPs during anticipation, action preparation, and consummatory stages of reward processing, these findings are inconsistent due to temporal and spatial overlap between the relevant electrophysiological components. Our results show three components following cue presentation are sensitive to incentive cues (N1, P3a, P3b). In contrast to previous research, reward-related enhancement only occurred in the P3b, with earlier components more sensitive to break-even and loss cues. During feedback anticipation, we observed a lateralized centro-parietal negativity that was sensitive to response hand, but not cue type. We also show that use of PCA on ERPs reflecting reward consumption successfully separates the reward positivity from the independently modulated feedback-P3. Lastly, we observe for the first time a new reward consumption component: a late negativity distributed over the left frontal pole. This component appears to be sensitive to response hand, especially in the context of monetary gain. These results illustrate that the time-course and sensitivities of electrophysiological activity that follows incentive cues do not follow a simple heuristic in which reward incentive cues produce enhanced activity at all stages and sub-stages.

**Keywords**: Reward, ERP, PCA, P3, N1, Reward Positivity, MID

**Author Notes**

Portions of this work were presented at the 55th annual meeting of the Society for Psychophysiology.

**Acknowledgements**

Portions of this work were funded by Australian Research Council Discovery Project DP150104514.

# 1. Introduction

Two vital aspects of reward processing, the anticipatory and consummatory stages, have been frequently investigated using variations of the monetary incentive delay (MID) task (Knutson, Westdorp, Kaiser, & Hommer, 2000). Each experimental trial in the MID task begins with the presentation of an incentive cue indicating the possible outcome for that trial: reward, loss, or break-even. After a brief anticipatory period, a target stimulus is presented that requires a behavioural response. If participants respond successfully to the target they are rewarded with money; otherwise they break-even or are penalized.

Research using fMRI has found that anticipation of rewards and losses in the MID task are associated with an increased BOLD signal in the dorsal anterior cingulate (dACC), and supplementary motor region (Knutson & Greer, 2008; Knutson et al., 2000). Anticipation of reward has also been associated with a selective increase in activity in the nucleus accumbens (NAcc; Knutson, Taylor, Kaufman, Peterson, & Glover, 2005), a striatal structure implicated in assigning values to predictive stimuli (Berridge, 2007; Berridge, Robinson, & Aldridge, 2009; Haber & Knutson, 2009). Reward consumption, on the other hand, has been associated with activity in ventromedial frontal cortex structures (Knutson, Fong, Adams, Varner, & Hommer, 2001).

Understanding reward anticipation and consumption requires not just the assessment of spatial anatomical correlates revealed by fMRI but also its temporal dynamics using event-related potentials (ERPs). While there is extensive research on ERPs associated with reward consumption and feedback processing (Bellebaum & Daum, 2008; Bellebaum, Kobza, Thiele, & Daum, 2010; Foti, Weinberg, Bernat, & Proudfit, 2015; Hajcak, Moser, Holroyd, & Simons, 2006; Proudfit, 2014), very few studies have investigated ERPs associated with reward anticipation *and* consumption in MID-type tasks. Those that have present inconsistent findings with respect to reward- and loss-related ERP modulation. As Novak and Foti (2015) have pointed out, electrophysiological components important for reward anticipation and consumption are susceptible to spatial and temporal overlap. This limitation could explain why consistent results in the context of MID-type tasks continue to remain elusive.

## 1.1. Reward Cue

While only a small number of studies have investigated ERPs in the MID task, a wide range of components are reported to be sensitive to reward, loss, and break-even cues. Given the motivational and affective significance of these cues as secondary reinforcers, their capacity to modulate ERPs is to be expected. What is unclear is which components reflect *functionally*

important processes in response to incentive cues in the MID task. It is important to note that the surface deflections we observe and that constitute ERPs are not in themselves components. Moreover, components are not isomorphic to discrete brain processes. Rather, ERPs and components that are assigned particular labels (e.g., N1, P3) reflect systematic variation in neural activity (Luck, 2014).

Several studies have reported that early components associated with low-level vision, post-perceptual processing, cognitive control, and template updating are sensitive to reward and loss cues relative to break-even cues. Results, however, are inconsistent, with null and entirely opposite effects commonly reported. For example, greater N1 amplitudes have been observed following reward cues relative to break-even cues in some studies (Doñamayor, Schoenfeld, & Münte, 2012), while other studies have observed no effect (e.g., Goldstein et al., 2006; Santesso et al., 2012). Greater P2 amplitudes have been observed for loss cues relative to reward (Santesso et al., 2012), but not in all studies (Goldstein et al., 2006). Modulation of N2 by cue type provides a particularly salient illustration of this inconsistency. While Novak and Foti (2015) observed greater N2 amplitudes for reward cues relative to loss and break-even cues, others have observed greater amplitudes for break-even cues relative to reward cues (Doñamayor et al., 2012; Pornpattananangkul & Nusslock, 2015) and for loss cues relative to reward cues (Yu & Zhou, 2006).

P3 effects in the context of the MID have been interpreted to reflect attentional allocation to motivational stimuli, irrespective of whether they are appetitive or aversive (Novak & Foti, 2015; Pfabigan et al., 2014). Again the results are inconsistent suggesting that processes associated with the P3 may not be consistent across studies. Greater P3 amplitudes have been observed in response to: (1) reward cues relative to break-even cues (Broyd et al., 2012; Goldstein et al., 2006; Novak & Foti, 2015; Pfabigan et al., 2014; Pornpattananangkul & Nusslock, 2015; Vignapiano et al.); (2) loss cues relative to break-even cues (Broyd et al., 2012 for null result; Novak & Foti, 2015; Vignapiano et al.); (3) reward cues relative to loss cues (Broyd et al., 2012 for null result; Novak & Foti, 2015; Pfabigan et al., 2014; Santesso et al., 2012); and (4) loss cues relative to reward cues (Vignapiano et al.).

ERPs involved in action anticipation and preparation have also been inconsistent. Some studies have observed a greater contingent negative variation (CNV) following reward and loss cues relative to break-even cues (Novak & Foti, 2015; Plichta et al., 2013), but once again other studies have not (Goldstein et al., 2006; Sobotka, Davidson, & Senulis, 1992; Vignapiano et al.). If there is reward-related modulation of the CNV it appears to be restricted to late-stages, 200ms preceding behavioural response. Research also suggests the enhancement of preparatory activity

for self-timed actions. Pornpattananangkul and Nusslock, 2015 observed a significantly larger readiness potential (RP) when participants were preparing to respond to a reward-cued temporal-bisection task.

### 1.2. Target

The prospect of receiving reward has been found to modulate ERPs associated with the presentation of target stimuli in MID tasks. Recently the target-P3 was found to be greater to target stimuli following reward cues (Broyd et al., 2012), consistent with earlier research showing greater P3 amplitudes to target stimuli associated with receiving reward (Homberg, Grunewald, & Grunewald-Zuberbier, 1981; Otten, Gaillard, & Wientjes, 1995; Ramsey & Finn, 1997). Moreover, the target-P3 has also been found to be enhanced for target stimuli following loss cues (Broyd et al., 2012).

### 1.3. Anticipation of Results

Incentive cue effects can also occur when anticipating feedback stimuli. Stimulus preceding negativity (SPN) amplitudes are modulated whenever a reward or punishment is possible. The SPN presents as a sustained slow negativity that is most pronounced over the right hemisphere and has been found to be reliably enhanced by reward and punishment expectancies. Across several studies, larger SPN amplitudes have been observed when participants are anticipating feedback that indicates they will receive a reward or avoid punishment, relative to when they are anticipating feedback that has no extrinsic motivational value (e.g., they can be neither rewarded or punished; Brunia, Hackley, van Boxtel, Kotani, & Ohgami, 2011; Ohgami, Kotani, Hiraku, Aihara, & Ishii, 2004; Ohgami et al., 2006).

In one study, the SPN has also been found to be modulated as a function of incentive cue in a variation of the MID task, with larger SPN amplitudes on trials where a reward is possible relative to those where no reward is possible (Pornpattananangkul & Nusslock, 2015).

### 1.4. Consumption

MID tasks have operationalized reward consumption through presenting feedback stimuli. Feedback stimuli indicate the accuracy and consequently the outcome of target detection and self-paced actions. Although this use of the term 'reward consumption' is prevalent in the reinforcement learning and MID literature (Baskin-Sommers & Foti, 2015), reward consumption likely encompasses additional processes that *follow* the receipt of visual or auditory feedback, and constitute the delivery and consumption of an actual physical reward (e.g., sucrose solution in animal studies; Castro & Berridge, 2014). While the issue of which ERPs reflect these different aspects of reward consumption is important, it is outside the scope of the present study, which is

restricted to feedback-locked ERPs. Within this class of stimulus, reward consumption has been associated with two distinct ERPs: Reward positivity (RewP) and feedback-P3 (fb-P3).

The RewP is positive frontocentral deflection occurring 200 to 300ms that is strongest following reward feedback (Bellebaum, Polezzi, & Daum, 2010; Hajcak et al., 2006; Proudfit, 2014). Multiple neuronal generators are associated with the RewP, including the ACC (Gehring & Willoughby, 2002; Hauser et al., 2014) and regions implicated in reward processing (Haber & Knutson, 2009) such as the ventral striatum and subregions of the medial PFC (Becker, Nitsch, Miltner, & Straube, 2014; Carlson, Foti, Harmon-Jones, & Proudfit, 2015; Carlson, Foti, Mujica-Parodi, Harmon-Jones, & Hajcak, 2011).

The RewP is also referred to as the Feedback Related Negativity (FRN). From a measurement standpoint, one is simply the inverse transformation of the other. Although the RewP and FRN are argued to be synonymous (Proudfit, 2014), with the core measurement difference being their polarity in difference waves, there are differences in interpretation. The FRN research literature has typically focused on the reinforcement learning functions that the component reflects, with dominant theoretical models proposing that it signifies either absolute prediction errors (e.g., the outcome differs from expected), or reward prediction errors (e.g., the outcome is worse than expected) specifically (Holroyd & Yeung, 2012). These models have received broad support from a range of empirical studies showing that FRN (or RewP) amplitudes are larger when outcomes are unexpectedly better or worse (e.g., Holroyd, Krigolson, & Lee, 2011; Ichikawa, Siegle, Dombrovski, & Ohira, 2010; Pfabigan, Alexopoulos, Bauer, & Sailer, 2011).

Despite its ubiquitous presentation in equiprobable gambling tasks (e.g., Angus, Kemkes, Schutter, & Harmon-Jones, 2015; Foti & Hajcak, 2010; Hajcak et al., 2006; Hajcak, Moser, Holroyd, & Simons, 2007; Holroyd, Krigolson, & Lee, 2011) and expectancy violation tasks (Holroyd, Krigolson, & Lee, 2011; Ichikawa, Siegle, Dombrovski, & Ohira, 2010; Pfabigan, Alexopoulos, Bauer, & Sailer, 2011), only four MID studies have reliably observed greater RewP amplitudes following reward feedback than following loss or break-even feedback (Novak & Foti, 2015; Pfabigan et al., 2015; Santesso et al., 2012; Yu & Zhou, 2006). Other MID studies have either observed RewPs in response to reward, loss and break-even feedback (Broyd et al., 2012; Pornpattananangkul & Nusslock, 2015) or no clear RewP at all (Doñamayor et al., 2012).

Overlapping with the RewP is the fb-P3, which is sensitive to expectancy violation (Bellebaum & Daum, 2008) and reward magnitude (Yeung & Sanfey, 2004). Greater fb-P3 amplitudes have been observed in MID tasks following feedback when participants expected to win or lose money relative to expecting to break even (Doñamayor et al., 2012; Novak & Foti,

2015; Pornpattananangkul & Nusslock, 2015). The fb-P3 has also been observed when participants successfully won or avoided losing money relative to only receiving performance feedback (Broyd et al., 2012)

Lastly, the presentation of feedback stimuli in the MID task has occasionally been found to modulate centro-parietal slow waves such the late positive potential (LPP), with enhanced amplitudes in response to feedback indicating poor performance regardless of whether this resulted in financial loss or not (Pornpattananangkul & Nusslock, 2015).

### 1.5. The Present Study

While a variety of electrophysiological components are elicited by the MID task, the effect of reward and loss anticipation and consumption on these components has been inconsistent. A likely explanation for this inconsistency is the spatial and temporal overlap of electrophysiological components. For example, the substantial scalp and temporal distribution of the P3 means it overlaps with other ERP components modulated by motivational and affective variables such as the Early Posterior Negativity and P3a. Similarly, the RewP and fb-P3 are present over similar timescales making the identification and quantification of one or both components difficult (Novak & Foti, 2015). The spatial and temporal overlap of ERPs confounds the measurement and interpretation of ERP results obtained in the MID task. Although several previous studies (Broyd et al., 2012; Novak & Foti, 2015; Pornpattananangkul & Nusslock, 2015) have examined ERPs that occur during the MID task, these studies have used quantification approaches that require *a priori* assumptions about which electrocortical activity will reflect functionally important processes. Only one study (Doñamayor et al., 2012) has used data driven exploratory approaches, and the permutation methods used in that study do not allow for the decomposition of overlapping components.

To overcome these issues and accurately assess what electrophysiological activity reflects *functionally* important processes during reward anticipation and consumption in the MID task, we used exploratory temporospatial principle components analysis (PCA) to quantify ERPs. Specifically, we applied PCA to ERPs generated at each stage of the MID task: cue, target, response preparation, and feedback. Temporospatial PCA has several advantages over typical approaches to ERP quantification. Firstly, temporospatial PCA is well suited to exploratory analysis of ERPs as it allows for the extraction of variance across time (e.g., samples) and space (e.g., electrodes), which decomposes an ERP waveform into its constituent components. This decomposition approach does not require *a priori* assumptions regarding the specific time-points or location that measurements will be taken from. Secondly, temporospatial PCA allows for the

identification of components that are obfuscated by temporal and spatial overlap, and prove difficult to measure using traditional ERP analyses (Dien & Frishkoff, 2005).

Although this study was exploratory, we did have general predictions regarding which factor combinations would likely be identified as functionally relevant. Firstly, we anticipated that factor combinations reflecting the P3 and CNV following incentive cues would be modulated by cue type, with greater amplitudes for gain rather than loss or break-even cues. Secondly, we anticipated that a factor combination reflecting the readiness potential preceding responses would be modulated by response hand and incentive cue, such that amplitudes would be larger for the contra-lateral hand and following gain incentives than following loss or break-even. Thirdly, we anticipated that a factor combination reflecting the SPN prior to feedback would be enhanced by incentive cues related to expected gains or losses. Lastly, we anticipated that the factor combinations that reflect the RewP and fb-P3 following feedback presentation would be enhanced for trials where participants received positive rather than negative feedback.

## 2. Method

### 2.1. Participants

20 males (Mean age = 23.6, *SD* = 4.0) participated in exchange for monetary compensation (AU$15 per/hr). Participants also received an additional AU $15 'bonus money' at the end of the experiment. All participants were right handed, and without past or current diagnoses of an affective disorder. This study was approved by the University of Sydney Human Research Ethics Committee and all participants provided informed written consent. Electrophysiological data from one participant was excluded from analyses due to insufficient usable trials.

### 2.2. Procedure

Upon arrival participants were provided with a brief overview of the study and consent form. Participants were not informed they could receive 'bonus money' until after they had provided written consent. Participants then completed a brief demographic questionnaire. Electroencephalography (EEG) and electrooculogram (EOG) electrodes were attached to the participant and they were seated in a darkened soundproofed testing chamber. The MID task was explained to the participant and they were guided through at least 10 practice trials. Once the experimenter was satisfied the participant understood the task, the experiment began. At the end of the task, participants completed a series of post-task questions. After these questions were completed the electrodes were removed and the participant was debriefed. Experimental materials were displayed on a 24-inch LCD computer screen with a refresh rate of 100 Hz.

Stimulus display and timings were controlled using the Psychophysics Toolbox for MATLAB (version 2013b).

**2.3. Design**

The task consisted of 360 trials divided into two 180 trial blocks. Each block was comprised of 60 'probable gain', 60 'probable loss', and 60 'break-even' trials. Each trial consisted of four key stages designed to produce different facets of reward and nonreward related processes. First, to examine reward anticipation related processes, we presented participants with an incentive cue. Then, to examine incentive related modulation of target processing, we presented participants with a target stimulus. To examine response preparation modulation, we asked participants to complete a simple temporal bisection task. Participants were instructed that to complete the bisection task, they had press a key when they thought 1000ms had elapsed from the appearance of the target. Participants were informed that successful performance on of the temporal bisection task would increase the probability of winning money rather than losing money. Finally, to examine consummatory processes, we presented participants with feedback regarding their performance.

Each trial began with a fixation-cross in the centre of the screen. After a variable interval of approximately 1200ms, the fixation cross was replaced with a leftward or rightward pointing arrow for 200ms, serving as the incentive cue. On probable gain, probable loss, and break-even trials, the arrow was green, red, or black, respectively. Participants were informed of the contingency between the arrow colour and probable outcome. Because we wished to tease apart activity related to motor preparation from activity related to incentive cue type, we instructed participants to use either their left or right hand on the temporal bisection. Moreover, if there were an interaction between these response hand and cue type, it would suggest that the incentive based modulation of electrocortical activity involved in, for example, the potentiation of response preparation is specific to response hand.

The direction of the arrow indicated which hand participants would need to use on the temporal bisection task on that trial. When the arrow (regardless of colour) faced left, participants had to press the 'z' key using the index finger of their left-hand. When the arrow faced right, participants had to press the '/' key with the index finger of their right-hand. Probable Gain and Probable Loss trials were mapped to opposite hands within each block, with the hand used alternating between blocks (e.g., left-hand response on probable gain, right-hand response on probable loss). Break-even trials were made with the left or right hand, in equal proportions within each block.

In order to distinguish activity associated with anticipation of the target stimulus from activity associated with motor preparation, we embedded a temporal bisection task within the MID task (see Pornpattananangkul & Nusslock, 2015). In this version of the MID, the target stimulus indicates the point from which participants are required to judge 1000ms passing. A black square was used as the target stimulus. The target stimulus was presented 2000ms after the onset of an arrow and remained on screen for 200ms.

Lastly, the outcome of each trial was signalled by feedback stimulus presented 1000ms after a response, or after 4000ms had passed from target onset (i.e., the latest possible time a participant would be allowed to make a response on that trial). Feedback stimulus remained on screen for 1000ms. Gains were signalled by a green upward pointing arrow and losses by a red downward pointing arrow. On break-even trials, an equals sign was presented, even when an incorrect response or no response was made. Probable gain and probable loss trials were further subdivided on the basis of the actual outcome. On 39 trials, outcomes were congruent (e.g., winning money on probable gain trial) and on the remaining 21 trials outcomes were incongruent (e.g., losing money probable gain trial). The actual response made by participants only affected the outcome of Probable Gain and Probable Loss trials when: 1) they made a response with the incorrect hand (e.g., left hand response on a right hand response trial), 2) they responded too quickly, which was defined as 250ms following the presentation of the target stimulus, or 3) they failed to make a response. We chose to do this as we wished to have a fixed ratio of trials in which participants received win/loss feedback. Because performance could have varied between expected win and expected loss conditions, we were concerned that this performance mismatch could have led to systematic imbalances in the number of trials available for feedback-locked analysis. An example trial is depicted in Figure 1.

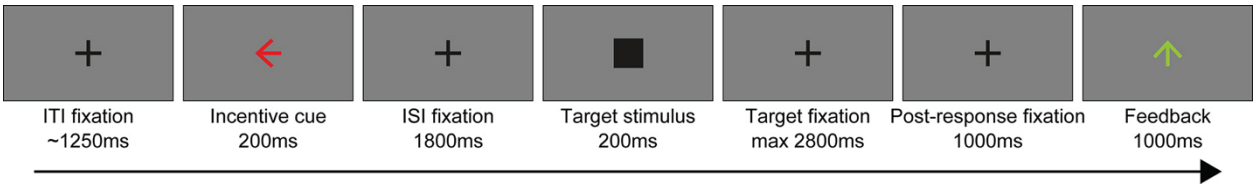


**Figure 1.** In this trial, the participant was required to form the intention to press the "z" key using their left hand, doing so in a context where they will probably lose – rather than gain – money. Participants were instructed that executing this action as close to 1000ms following the onset of the target stimulus increased the likelihood of winning money rather than losing it. At the end of the trial, the participant receives feedback signalling that they had won money.

Participants were informed that they start on $10 and that on every successful gain-outcome trial they would gain $0.25, while on every loss-outcome trial they would lose $0.25.

### 2.4. Physiological Recording & Data Reduction

EEG was recorded from 60 active Ag/AgCl electrodes (ActiCap, Brain Products, Gilching, Germany) located according to the International 10-20 system. EOG was recorded from two electrodes placed 1cm lateral to the outer canthi of each eye (for horizontal EOG) and from one electrode placed on the sub-orbital region of the right eye (for vertical EOG, e.g., Hofmann, Kuchinke, Tamm, Võ, & Jacobs, 2009). Raw data was amplified and sampled at 1000Hz by a Brain Products QuickAmp72, referenced to a common average with a ground electrode located at AFz and recorded using Brain Vision Recorder (version 1.20, Brain Products, Gilching, Germany).

Raw EEG data was preprocessed offline using native EEGLAB (Delorme & Makeig, 2004) and ERPLAB (Lopez-Calderon & Luck, 2014) functions. Continuous EEG data were down sampled to 256Hz, re-referenced to the average of TP9 and TP10, bandpass filtered between .01-30hz (IIR Butterworth, 25db roll-off), and residual line noise was reduced using a 50Hz Parks-McClellen notch filter. Sections of continuous data containing gross movement artefacts were removed and Independent Components Analysis (ICA) was used to remove electrophysiological artefacts introduced by eye movements (Jung et al., 2000).

Separate segments were created for each 'stage' of reward processing. For incentive cue segments the period 200ms prior through to 2000ms after cue presentation was selected. For target segments the period 200ms prior through to 500ms after the target presentation was selected.[83] For response segments the period 1000ms prior through to 200ms after participant's responses was selected. To examine responses that occur in anticipation of feedback, a period 1000ms prior to feedback presentation was selected. Finally, feedback segments were created using the period 200ms prior through to the 1000ms after feedback presentation. Cue, target, and feedback segments were baseline corrected to their 200ms pre-stimulus periods. Response-locked activity was baseline corrected to the period 200ms prior to target stimulus presentation, while feedback anticipation activity was baseline corrected to the period 200ms following responses. Segments from trials in which participants made an incorrect response (e.g., used the incorrect key, failed to respond, or responded in < 250ms) were excluded from analyses. Retained segments were submitted to artefact rejection processes implemented in ERPLAB (Lopez-Calderon & Luck, 2014). Segments were rejected if on any electrode 1) the voltage difference was greater than 200uv, 2) the voltage changed by more than 50uv within a 200ms

[83]Although 16 factor combinations each accounted for greater than 1% of variance in target ERPs, and several of these factor combinations appeared to reflect ubiquitous components previously found to be effected by reward and loss incentives (e.g., the P3), these were not significantly modulated in the current task by cue type (all $p > .047$), response hand (all $p > .023$), or an interaction between the two (all $p > .072$; adjusted critical $p = 0.003125$).

sliding window across the entire segment, or 3) the voltage changed by more than 50uv between samples.

Subject averages for cue, target, response and feedback anticipation ERPs were binned according to the cue type (probable gain, probable loss, break-even) and response hand (left, right). Feedback activity was binned according to according to cue type (probable gain, probable loss), response hand (left, right), and feedback type (positive, negative). Break-Even trials were not included in the analysis of feedback ERPs as only a single outcome was ever possible. Participants were excluded from the analysis of a task stage if there were not at least 10 trials in the bin. One participant was excluded from feedback analysis due to having 9 useable trials in one bin. Descriptive statistics for available and retained trials are presented in Table 1. Grand averages for all bins and segment types are presented in supplementary figures 1 (Cue), 2 (Target), 3 (Response) and 4 (Feedback anticipation) and5 (Feedback).

**Table 1.**

*Mean Trials Retained for Analysis in Each Condition*

| | Cue | Response | Feedback Anticipation | Feedback |
|---|---|---|---|---|
| Probable Gain, left hand | 53.6 | 55.3 | 55.7 | |
| Probable Gain, right hand | 52.8 | 54.4 | 55.2 | |
| Probable Loss, left hand | 52.4 | 53.2 | 54.3 | |
| Probable Loss, right hand | 52.7 | 54.5 | 55.5 | |
| Break-Even, left hand | 52.4 | 53.0 | 54.1 | |
| Break-Even, right hand | 51.0 | 52.2 | 54.0 | |
| Probable Gain, left hand, win | | | | 36.2 |
| Probable Gain, left hand, loss | | | | 19.6 |
| Probable Gain, right hand, win | | | | 36.1 |
| Probable Gain, right hand, loss | | | | 18.8 |
| Probable Loss, left hand, win | | | | 19.1 |
| Probable Loss, left hand, loss | | | | 35.7 |
| Probable Loss, right hand, win | | | | 20.1 |
| Probable Loss, right hand, loss | | | | 36.3 |

### 2.5. Statistical Analysis

Quantification of ERP components was conducted using the PCA Toolkit (Dien, 2010a). PCA is a factor analytic method that identifies and separates linear combinations of data points across temporal and spatial domains, allowing identification of overlapping electrocortical activity. All PCA analyses followed an identical process. First, temporal PCA was applied using samples as variables, and conditions, participants, and electrodes as observations. Consistent with previous recommendations, we used a Promax rotation with Kaiser normalization (Dien, 2010b). The number of temporal factors extracted for rotation in each PCA was determined using the parallel

test (Cattell, 1966). Second, temporal factors were submitted to a spatial PCA using Infomax rotation (Dien, 2010b; Dien, Khoe, & Mangun, 2007) with electrodes as variables, and conditions, participants, and temporal factor loadings as observations. Factors extracted for rotation were again determined using the parallel test (Cattell, 1966). Grand averages of example data submitted to PCA are presented in Figure 2. More comprehensive grand averages for each stage of reward processing are included in supplementary information 1, 2, 3, 4 and 5.

To aid interpretation, peak factor loadings were converted into microvolts. For all task stages, factor combinations were retained for subsequent analyses if they accounted for more than 1% of variance (e.g., Foti, Hajcak, & Dien, 2009). For each set of analyses, we report the total number of extracted factors and total number of retained factors that met the 1% threshold. The latter were subjected to robust Analysis of Variance (ANOVA) in accordance with published recommendations (Dien, Franklin, & May, 2006; Keselman, Wilcox, & Lix, 2003), and as implemented in the PCA Toolkit (Dien, 2010a). This approach to null hypothesis testing has been shown to be robust to violations of homogeneity and non-normal distributions, and reduces Type 1 error rate (Keselman et al., 2003). Each ANOVA used a starting seed of 100, was bootstrapped 5000 times, and 5% upper and lower means trimming. Robust ANOVA tests are indicated by "$T_{WJt/c}$", and the interpretation of this statistic and resulting *p* values are identical to a conventional ANOVA. Significance thresholds for omnibus tests were corrected for multiple comparisons using the Bonferroni method, with the number of extracted factors subjected to the ANOVA determining the magnitude of the correction. Significance thresholds for pairwise comparisons were also corrected for multiple comparisons.

Regarding behavioural data, mean response times on correct trials (e.g., participants pressed the instructed key 250-3000ms following target onset) first underwent a linear transform, where 1000ms was subtracted from each response, providing a more easily interpretable measure of performance. The resulting values were then subjected to a repeated measures ANOVA as implemented in SPSS (IBM, V20), with partial eta squared is provided as a measure of effect size.

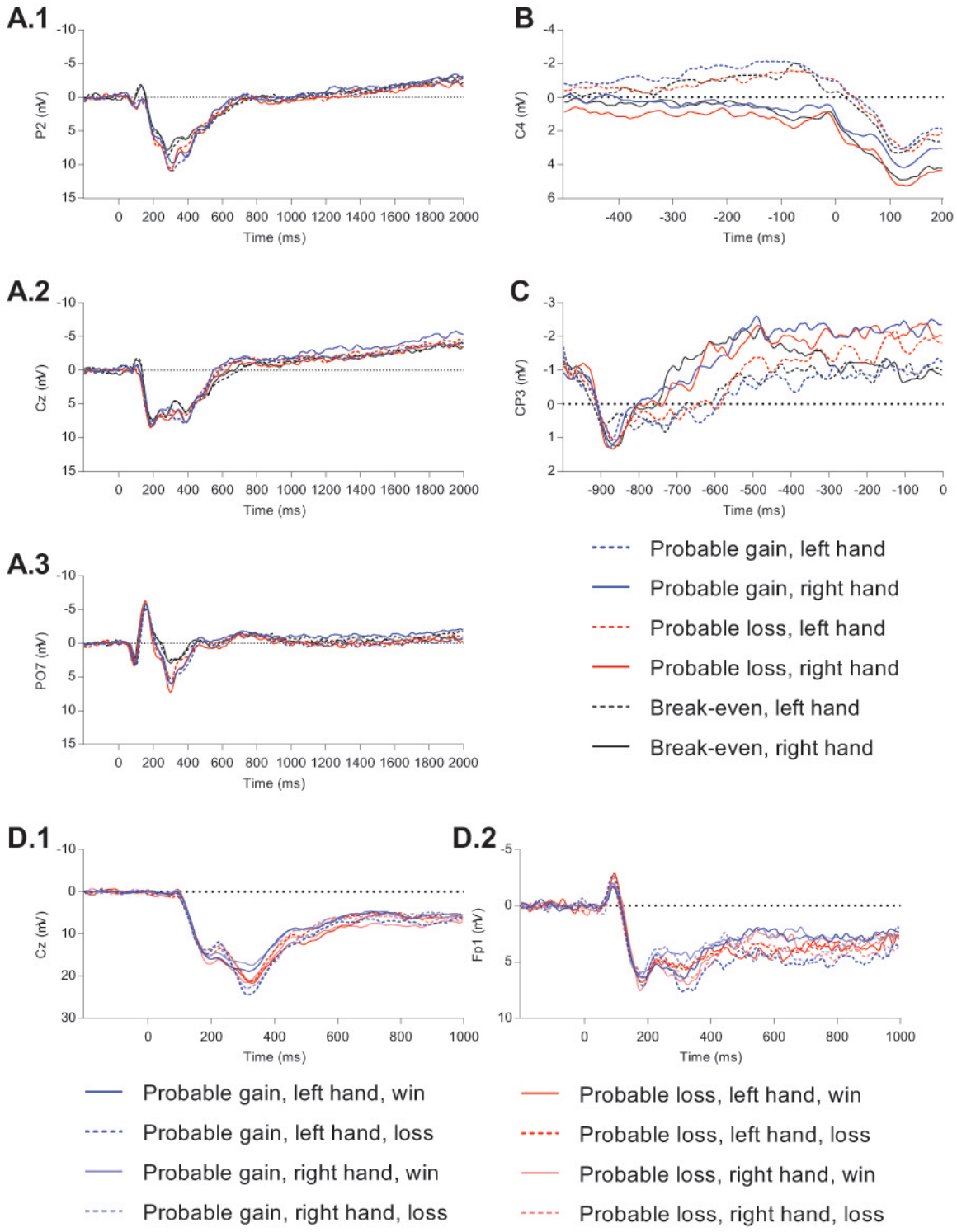


**Figure 2.** Grand average ERPs prior to PCA for cue-locked (A.1, A.2, A.3), response-locked (B), feedback anticipation (C), and feedback-locked (D.1, D.2) activity. Sites presented were those identified as being the spatial peak of relevant temporospatial factor combinations.

## 3. Results

### 3.1. Behavioural Data

Response time values were submitted to a three (cue type: probable gain, probable loss, break-even) x two (response hand: left, right) factor ANOVA. Consistent with previous research indicating improved accuracy on incentivized trials, there was a main effect of cue type ($F(2,36) = 3.65$, $p = .036$, $\eta_p^2 = .17$). Post-hoc tests indicated that this effect was driven by a more accurate response times on probable gain trials ($M = 232.39$, $SD = 402.52$) than on break-even trials ($M = 262.78$, $SD = 412.21$; $p = .018$). There was no difference between probable loss trials ($M = 241.30$, $SD = 410.29$) and probable gain trials ($p = 1.00$) or break-even trials ($p = .377$). There was no effect of response hand on response times ($F(1,18) = 1.57$, $p = .226$, $\eta_p^2 = .08$), or an interaction between incentive cue and response hand ($F(2,36) = .81$, $p = .398$, $\eta_p^2 = .043$).

### 3.2. Cue Activity

A total of 18 temporal and 4 spatial factors were extracted for cue-locked ERPs. Of these 72 factor combinations, 16 explained at least 1% of variance and were submitted to a three (cue type: probable gain, probable loss, break-even) x two (response hand: left, right) factor ANOVA with a Bonferroni corrected significance threshold of $p < .0031$. Three factor combinations were sensitive to cue type; their waveforms and scalp maps are presented in Figure 3.

The earliest factor combination, TF8/SF1 (Temporal Factor 8/Spatial Factor 1), was maximal over P2 between 132-136ms, and appears to reflect a parietal N1, $T_{WJt/c}(2.0,16.0) = 9.85$, $p = 0.0014$. Pairwise comparisons between cue types indicate the negativity is stronger for break-even cues than probable gain cues ($T_{WJt/c}(1.0,18.0) = 7.28$, $p = 0.015$) and probable loss cues ($T_{WJt}/c(1.0,18.0) = 20.41$, $p=0.00020$). Factor loadings did not differ significantly between probable gain and probable loss cues, $T_{WJt/c}(1.0,18.0) = 0.50$, $p = 0.50$.

The next factor combination to be affected by cue type, TF6/SF1, had a scalp distribution and latency similar to the P3a and was maximal over Cz between 276-280ms, $T_{WJt/c}(2.0,16.0) = 15.34$, $p = 0.00060$. Follow up pairwise comparisons indicate that the loadings for probable gain cue were significantly smaller than probable loss ($T_{WJt/c}(1.0,18.0) = 22.67$, $p = 0.00080$) and break-even cues ($T_{WJt/c}(1.0,18.0) = 21.31$, $p = 0.00020$), which did not differ significantly from one another ($T_{WJt/c}(1.0,18.0) = 0.89$, $p = 0.36$).

Lastly, TF2/SF2, a P3b like factor combination, was maximal over PO7 between 360-364ms ($T_{WJt/c}(2.0,16.0) = 20.95$, $p = 0.00020$. Consistent with previous research indicating enhanced P3 amplitudes to incentive cues (Novak & Foti, 2015), factor loadings for TF2/SF2 were significantly greater for probable gain ($T_{WJt}/c(1.0,18.0) = 40.73$, $p<0.00000001$) and probable loss cues ($T_{WJt}/c(1.0,18.0) = 14.48$, $p=0.0020$) than break-even cues. TF2/SF2

amplitudes were also significantly greater for probable gain cues than probable loss cues ($T_{WJt}/c(1.0,18.0) = 11.95$, $p = 0.0040$). Although one factor combination appeared to reflect the CNV, this was insensitive to any of the experimental manipulations (see appendix 1). No other factor combinations were significantly affected by Cue type, response hand, or their interactions.[84]

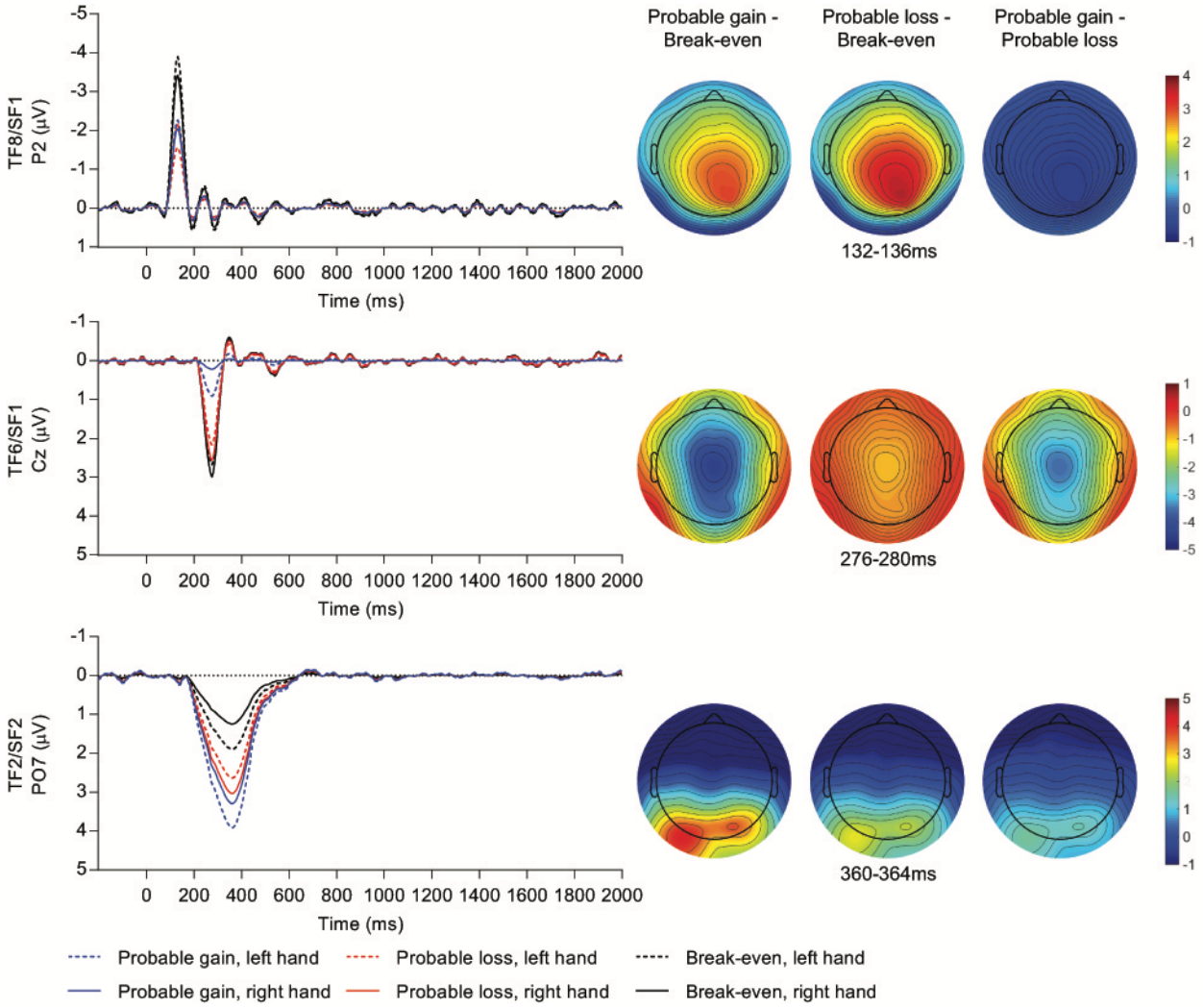


**Figure 3.** Cue locked waveforms and topographic maps for factor combinations reflecting the N1 (upper row), P3a (middle row), and P3b (bottom row). Waveforms and topographic maps are presented in microvolt scale. Topographic maps present the difference between cue types at the peak latency of each temporospatial factor combination.

[84]While there was reliable modulation of these factor combinations by Cue type, they were insensitive to response hand (TF8/SF1: $T_{WJt}/c(1.0,18.0) = 0.03$, $p = 0.87$, TF6/SF1: $T_{WJt}/c(1.0,18.0) = 0.00$, p = 0.99, TF2/SF2: $T_{WJt}/c(1.0,18.0) = 1.63$, $p = 0.23$), or the interaction between response hand and cue type (TF8/SF1: $T_{WJt}/c(2.0,16.0) = 3.66$, $p = 0.071$, TF6/SF1: $T_{WJt}/c(2.0,16.0) = 1.10$, $p = 0.39$, TF2/SF2: $T_{WJt}/c(2.0,16.0) = 1.85$, $p = 0.21$).

### 3.3. Response Activity

Response locked ERPs yielded 5 temporal and 4 spatial factors, with 17 of the 20 factor combinations accounting for at least 1% of variance. These factors combinations were submitted to a three (cue type: probable gain, probable loss, break-even) x two (response hand: left, right) factor ANOVA with a Bonferroni corrected threshold of $p < .0029$.

Only one factor combination, TF1/SF3, was sensitive to experimental manipulation. TF1/SF3 was a sustained right hemisphere negativity maximal over C4 132-128ms prior to responses and it likely reflects a readiness potential (RP). As shown in Figure 4, TF1/SF3 is modulated by response hand ($T_{WJt/c}(1.0,18.0) = 46.98$, $p < 0.00000001$). Consistent with basic research on motor preparation ERPs (Brunia, 1988), amplitudes were significantly more negative for left hand responses than right hand responses. Moreover, the direction of this effect was reversed when amplitudes were measured from C3.

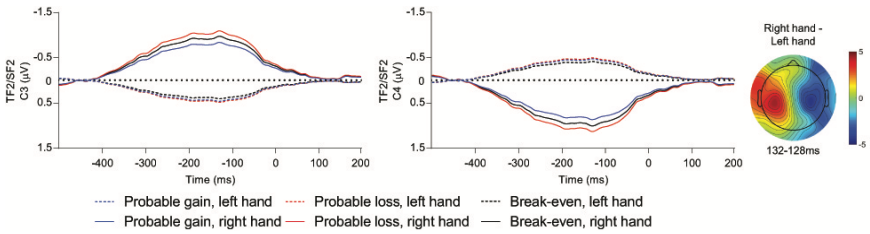


**Figure 4.** Response locked waveforms and topographic map for the factor combination that reflects the Readiness Potential. The waveforms and topographic map are presented in microvolt scale. The topographic map presents the difference between response hands at the peak latency of the temporospatial factor combination. Although this factor combination peaked over C4, the contra-lateral site, C3, is provided to illustrate the inversion of the waveform. Because this factor combination was maximal over C4 and for left-hand responses, the topographic map is most negative over the right hemisphere.

While TF1/SF3 appears to reflect a readiness potential, contrary to some recent research, it was not significantly affected by cue type ($T_{WJt}/c(2.0,16.0) = 0.75$, $p = 0.48$). Nor was there an interaction between cue type and response hand, $T_{WJt}/c(2.0,16.0) = 0.75$, $p = 0.48$.

### 3.4. Feedback Anticipation Activity

For activity that occurred prior to feedback presentation, 11 temporal factors were extracted, with 3 spatial factors each. 9 of these 33 factor combinations explained at least 1% of variance, and were subjected to a three (cue type: probable gain, probable loss, break-even) by two (response hand: left, right) factor ANOVA, with a Bonferroni corrected threshold of $p < .0056$. A single factor combination was sensitive to any of the experimental manipulations used in the present task, which is presented in Figure 5.

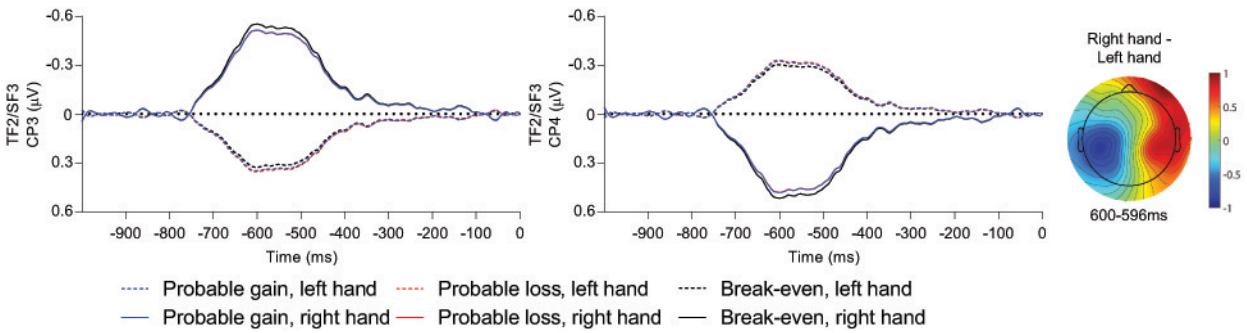


**Figure 5.** Feedback anticipation waveforms and topographic map for TF2/SF3. The waveforms and topographic map are presented in microvolt scale. The topographic map presents the difference between response hands at the peak latency of the temporospatial factor combination. Although this factor combination peaked over CP3, the contra-lateral site, CP4, is provided to illustrate the inversion of the waveform. Because this factor combination was maximal over CP3 and for right-hand responses, the topographic map is most negative over the left hemisphere.

This factor combination, TF2/SF3, was centroparietal negativity that peaked over CP3 600-596ms prior to the presentation of feedback. TF2/SF3 was reliably modulated by the response hand that participants had used, with more negative voltages over CP3 following right hand responses than left hand responses, $T_{WJt/c}(1.0,18.0) = 14.25$, $p = 0.0020$. As was also observed for the factor combination preceding actual motor responses (TF1/SF3), the direction of this effect was reversed when measured over CP4. Although the topology and sensitivity of this factor combination superficially resembles that of preparatory motor activity, it is important to emphasize that it occurred in anticipation of receiving feedback (i.e., 400-500ms *following* responses).

A factor combination that had a similar time-course and scalp topology to the SPN was also extracted, but was not affected by any of the experimental manipulations (see appendix 2).

### 3.5. Feedback Activity

For Feedback-locked averages, a total of 10 temporal factors and 4 spatial factors were extracted. Of these 40 factor combinations, 14 explained at least 1% of variance. These 14 factors were subjected to a two (cue type: probable gain, probable loss) x two (response hand: left, right) x two (outcome: gain, loss) factor ANOVA with a Bonferroni corrected threshold of $p < .0035$. This ANOVA yielded three factor combinations that were sensitive to different aspects of the task. The waveforms of which are displayed in Figure 6.

The earliest factor combination, TF4/SF1, was a frontocentral positivity that peaked over Cz between 224-228ms. TF4/SF1 was modulated by feedback outcome with more positive voltages observed in response to gain feedback than loss feedback, $T_{WJt/c}(1.0,17.0) = 15.20$, $p = 0.0034$. The topology and latency of this TF4/SF1 suggest it represents the RewP. This

replicates a previous application of PCA to reward-consumption ERPs (Foti & Hajcak, 2009) and confirming that the RewP can be observed in response to feedback stimuli in the MID task.

While TF4/SF1 was only modulated by outcome and not expected outcome or response hand, an overlapping factor, TF1/SF1, was sensitive to the interaction between cue type and feedback type, $T_{WJt/c}(1.0,17.0) = 22.82$, $p = 0.0014$. TF1/SF1 had a centroparietal distribution peaking over Cz, 380-384ms following feedback presentation, suggesting it corresponds to the fb-P3 observed in other studies (Balconi & Crivelli, 2010; Novak & Foti, 2015). Pairwise comparisons of the interaction between cue type and outcome, with a corrected significance threshold of $p < .0125$, revealed three significant contrasts. First, TF1/SF1 was significantly more positive on probable gain trials when participants received loss feedback relative to when they received gain feedback ($p = .0002$). Second, more positive amplitudes were observed in response to win feedback on probable loss trials relative to probable gain trials ($p = .007$). Third, the opposite pattern was observed in response to loss feedback, with significantly more positive amplitudes on probable gain trials relative to probable loss trials ($p = .009$). Taken together, these comparisons indicate that TF1/SF1 amplitudes were enhanced when the outcome of a given trial violated the expectation established by the cue.

Lastly, a factor combination, TF2/SF3, representing a late frontopolar negativity was modulated by the response hand used by participants on each trial, $T_{WJt}/c(1.0,17.0) = 22.06$, $p=0.0010$. TF2/SF3 peaked between 880-884ms following feedback over Fp1 and was more negative for right-hand than left-hand responses. There was also a significant interaction between response hand and feedback type, $T_{WJt}/c(1.0,17.0) = 12.28$, $p = 0.0022$. Pairwise comparisons indicate that gain feedback for right-hand responses elicited significantly more negative amplitudes than gain feedback for left-hand responses, $T_{WJt}/c(1.0,17.0) = 38.00$, $p < 0.00000001$. Amplitudes did not differ as a function of response hand for loss feedback, $T_{WJt}/c(1.0,17.0) = 0.19$, $p = .67$.

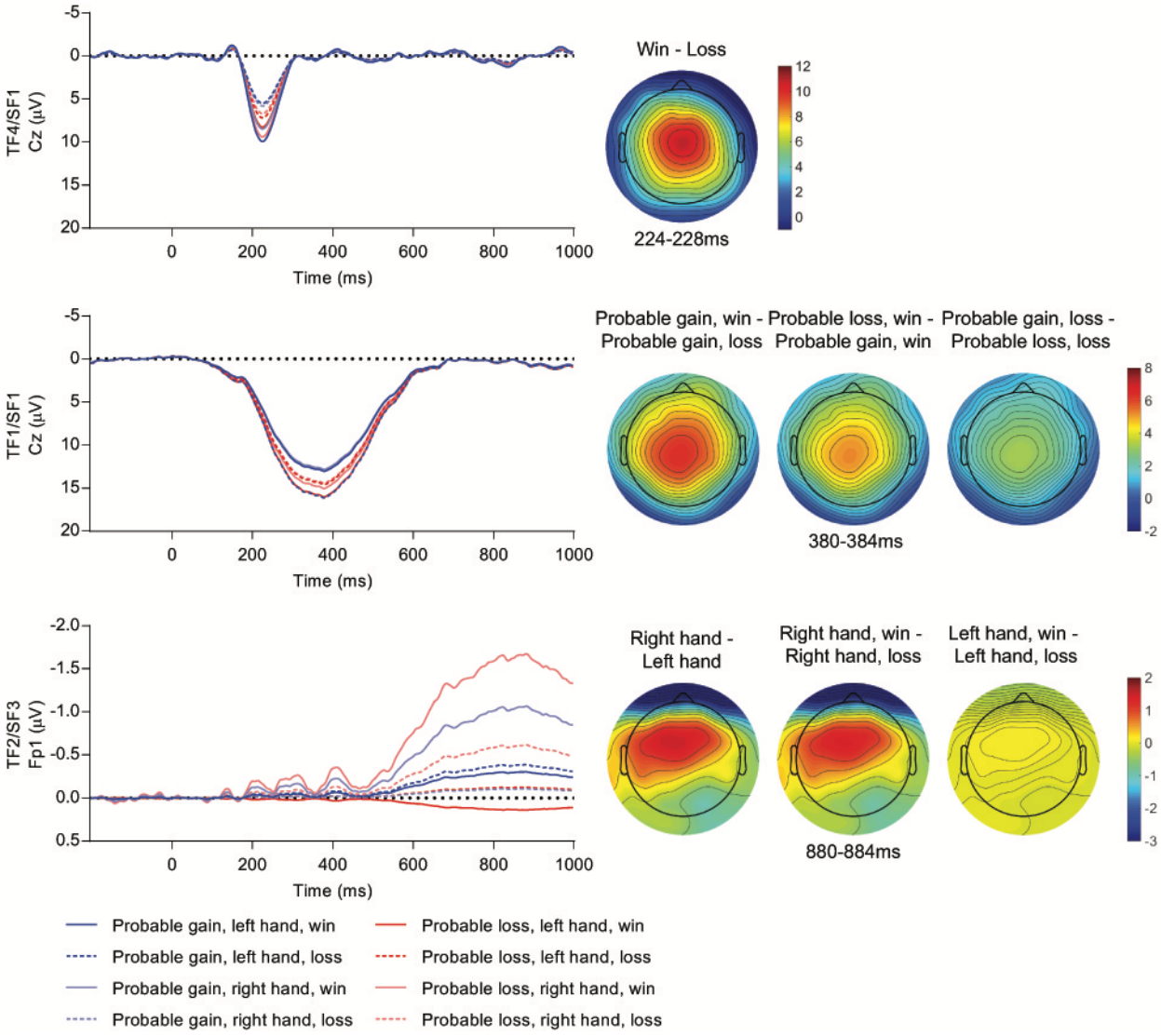


**Figure 6.** Feedback locked waveforms and topographic maps for factor combinations reflecting the Reward Positivity (upper row), feedback-P3 (middle row) and late frontopolar negativity (bottom row). Waveforms and topographic maps are presented in microvolt scale. Topographic maps present the difference between feedback valence (upper row), cue type and feedback valence (middle row), and feedback valence and response hand (bottom row) at the peak latency of each temporospatial factor combination.

## 4. Discussion

In this study we applied temporospatial PCA to data from a modified MID task. This novel approach was used in order to identify electrophysiological components important for reward anticipation and consumption in the MID task. PCA successfully extracted several components differentially modulated by stimuli cuing anticipation of probable gain, probable loss, and breaking-even. When applied to feedback anticipation, PCA was able to extract a factor combination that resembled the SPN. However, this factor combination was not reliably modulated by incentive cue or response hand. Instead, we observed a lateralized centroparietal

factor combination that was acutely sensitive to the hand participants had *just* used to make a response. The application of PCA to reward consumption also extracted well-established feedback processing ERPs in the RewP and fb-P3, components that have proved difficult to observe in the MID task. In addition, a new component was revealed in the late stages of feedback processing. This component is a late negativity distributed over the left frontal pole that appears to be sensitive to response hand, especially in the context of monetary gain. While a component reflecting the RP was identified preceding responses, it appeared to be insensitive to the prospect of gains, losses, and breaking-even. Target components as well appeared to be insensitive to cue type.

**4.1. Incentive Cues**

This study provides important insight into the modulation of ERPs by incentive cues in the MID task. Specifically, it provides evidence that the ERP components modulated by probable gain, probable loss, and break-even cues reflect distinct processes. There is no simple heuristic where reward incentive cues produce greater activity than loss and break-even cues. Furthermore, reward enhancement may not be present until relatively late in stimulus processing.

The early parietal factor combination, reflecting the N1, was enhanced for break-even cues relative to probable gain and probable loss. As the N1 is associated with early processes related to selective attention, this observation suggests 'neutral' incentive cues may receive some preferential processing relative to reward and loss cues. This finding is inconsistent with the results of one previous study where N1 amplitudes were enhanced for reward cues relative to break-even (Doñamayor et al., 2012). However, grand averages from other MID studies also appear to show a greater N1 to break-even cues, but these effects have not been directly reported (e.g., Study 1; Novak & Foti, 2015).

This early negativity was distinct from two subsequent, independently modulated centro-parietal and parietal positivities, which typically overlap to form the P3 (Polich, 2007). The first positivity had a latency and centro-pariatal distribution consistent with the P3a. The P3a is enhanced by stimuli requiring the automatic attentional switching or the initiation of inhibitory processes. Interestingly, P3a amplitudes were smaller for probable gain cues than probable loss or break-even cues, a finding broadly consistent with the observation that P3a amplitudes are enhanced by stimuli of negative affect (Delplanque, Silvert, Hot, Rigoulot, & Sequeira, 2006).

The second positivity reflects the P3b and has a predominantly parietal distribution. Consistent with previous research, P3b amplitudes were larger for probable gain cues relative to probable loss and break-even cues (Broyd et al., 2012; Goldstein et al., 2006; Novak & Foti, 2015; Pfabigan et al., 2014; Pornpattananangkul & Nusslock, 2015; Santesso et al., 2012). The

topology of the P3b that we observe was considerably left lateralized in contrast to the more central distributions observed in other studies. While it is difficult to identify a clear explanation for this lateralization, several studies have noted that object/color recognition and action processes may be more prevalent over the left hemisphere (Johnson-Frey, Newman-Norlund, & Grafton, 2005; Proverbio, Burco, del Zotto, & Zani, 2004). Aspects of these processes could conceivably be enhanced when an incentive cue indicates a possible reward. However, given the ubiquity of the P3b across multiple contexts, and considerable problems that reverse inference creates, we are reluctant to ascribe any particular psychological process as causative of this lateralization.

These data suggest that electrophysiological activity discriminates between different incentive cues and have distinct time-courses. First, an initial parietal negativity occurs that is more sensitive to break-even cues than probable gain or loss cues. Second, a centro-parietal positivity occurs that is more sensitive to probable loss and break-even cues than gain cues. Finally, a slow parietal positivity occurs that is more sensitive to probable gain cues than loss or break-even cues. Importantly, while these results replicate the common finding that the P3b/P3 is greater to reward incentive cues, it also suggests this enhancement is not present until relatively late in processing. In sum, while earlier components reflect capture by motivationally salient stimuli that signal increased probability of losses, the later positivity corresponding to the P3b/P3 reflects capture by motivationally salient stimuli that signal increased probability of gains.

In the context of incentive cues, earlier components (e.g., N1, P3a) may reflect the automatic capture and allocation of attention to stimuli signalling future events that may be disadvantageous: this includes both probable loss and break-even cues. Break-even cues require participants to expend physical and cognitive effort with no possibility of reward, representing a net-loss to participants. This conjecture is broadly supported by findings that non-reward feedback produce identical RewP amplitudes to loss feedback (Hajcak et al., 2006). This finding is also consistent with recent perspectives on the nature of incentive salience and incentive cue processing. Berridge and Kringelbach (2015) argue that even though an incentive cue may produce approach-related feelings of "wanting" in anticipation of a reward, this is not intrinsically a positive state. When the incentive cue – or features of the environment – simultaneously signals that a reward will not be forthcoming, this can produce a form of negative incentive salience, which is borne out as a feeling of frustration (Harmon-Jones, Harmon-Jones, & Price, 2013). Conversely, the presentation of a break-even cue could be also appraised as a relief from the possibility of loss that is associated with both probable loss and probable gain

trials. We speculate that whether break-even cue stimuli are interpreted as net losses or the relief from potential loss may be dependent on participants' perceptions that they can influence the outcome of each trial, and the development of expectations on this basis. Because blunted RewP amplitudes are observed when participants have limited perceived agency or sense of control over action-outcome contingencies (Angus et al., 2015; Bellebaum, Kobza, et al., 2010; Mühlberger, Angus, Jonas, Harmon-Jones, & Harmon-Jones, 2017), this speculation could be tested using experimental designs that directly manipulate perceived control, in conjunction with examination of RewP amplitudes.

In this study, we do not observe reliable incentive cue modulation of anticipatory slow-waves such as the CNV. These effects have proved difficult to measure in ERP versions of the MID task with inconsistency both in the presence of modulation and the effects of probable gain, loss and breaking-even cues. Nevertheless, recent research has reported a semi-consistent enhancement to later-stages of the CNV by reward cues relative to break-even cues (Novak & Foti, 2015; Pfabigan et al., 2014). While further research is required to confirm this effect, it is possible that methodological differences are responsible for our null effect. Typical MID tasks involve an incentive cue followed by a target stimulus signalling for an immediate response. The slow wave that precedes the target is therefore a combination of anticipatory processes (e.g., the early CNV) and action preparation (e.g., the late CNV). In our version of the task participants were not required to make a response to the target stimulus but instead begin the temporal bisection task. It is unsurprising then that a component involved in action preparation is unaffected.

**4.2. Responses**

Unlike previous research we failed to observe any enhancement of the RP in response to incentive cues (Pornpattananangkul & Nusslock, 2015). There are a number of possible reasons for this null effect. Firstly, when reported, the effect size of RP modulation - as with the CNV - by incentive cue is weak relative to the effect size of the anticipatory ERPs that follow cue presentation. While we may simply have required significantly more statistical power to detect these effects, it is possible these effects may not be robust.

Secondly, two key methodological differences may have influenced our result. First, unlike Pornpattananangkul and Nusslock (2015), the contingencies in our study were probabilistic. When a participant received a probable gain cue they were only more likely to gain money, whereas when they received a probable loss cue they were only more likely to lose. In contrast, in Pornpattananangkul and Nusslock (2015), incentive cues of various forms indicated with certainty what the outcome would be, provided that participants responses were accurate.

In our task, despite being told otherwise, the timing of participants' responses never actually guaranteed the reward outcome.[85] It is possible that RP modulation is sensitive to the efficacy of participant responses to secure the monetary gains or avoid the monetary losses indicated by the incentive cue. When accurate responses only *increase* rather than guarantee the possibility of reward, RP modulation may be reduced if not entirely absent.

Thirdly, it is also possible that while RP modulation is enhanced by incentive cues, it does not distinguish between motivations to secure monetary gain or avoid monetary loss. Unlike Pornpattananangkul and Nusslock (2015) who had an equal amount of reward and break-even trials, our study had an equal amount of probable gain, probable loss and break-even trials. The failure to observe RP modulation may due to the fact that the vast majority of our trials involved incentivised performance. Note that this could still occur even if the RP is not being modulated by the anticipatory cue as our results suggest. All that it requires is the experiment to be designed to emphasise incentivised performance.

**4.3. Feedback anticipation**

While we did observe a factor combination that reflects the SPN in topology and time-course, it was not modulated by any of the experimental manipulations. One reason for the absence of incentive cue modulation may be that participants' perception of control did not differ between Probable Gain, Probable Loss, and Neutral cue conditions. In the present study, the outcome of each trial was predetermined, and participants' responses could only alter this outcome if they failed to respond or responded using the incorrect hand. This element of the task design may have reduced participants' perceived control, and contributed to an absence of cue related modulation. Consistent with this explanation, research has shown that SPN amplitudes are smaller when participants believe they have limited control over outcomes than when they believe they have greater control (Mühlberger et al., 2017, also see Masaki, Yamazaki, & Hackley, 2010, Kotani, Ohgami, Yoshida, Kiryu, & Inoue, 2017).

We did, however, find a factor combination that was robustly modulated by the response hand that participants had just used. This factor combination was superficially similar to typical preparatory motor responses (e.g., RP, LRP) in its topology and lateralization as a function of response hand. Importantly, however, it occurred at least 400ms *after* a response had been made, presenting 600ms prior the receipt of feedback regarding the success or failure of that action. A conservative interpretation is that this factor combination reflects the residual motor activity that follows responses (e.g., a slow motor potential; Müller et al., 1994). The broader implications of this factor combination within the context of the MID task and reinforcement learning is

[85]Aside from failing to perform at all.

unclear, however we note that the particular context in which it has evoked is broadly consistent with the notion of an "eligibility trace" which is thought to produced following an action and serves as a form of working memory regarding the behaviour that will – or will not be – reinforced by feedback (Doya, 2008). Alternatively, this component may reflect activity associated with hand movements following participants' responses in the temporal bisection task. Specifically, following their response on each trial, participants likely returned their hand to a resting position, producing tactile and kinaesthetic activity. This tactile and kinaesthetic activity produces reafferent activity in the somatosensory cortex (Keysers, Kaas, & Gazzola, 2010). However, in the present study we did not collect surface electromyography (EMG) from participants' arms, and as such it is impossible to determine whether outcome anticipation or reafference better explains the results. Moreover, the omission of EMG measurements also limits our ability to address different explanations for CNV and RP modulation.

The limited modulation of SPN amplitudes by incentive cues and the presence of a post-motor component may also be due to the duration of the interval between motor responses and feedback onset. Previous studies have reported that post-movement activity can be observed contra-lateral to the response hand up to 2 seconds following responses (Damen & Brunia, 1994). Consequently, this activity may mask SPN modulation. However, that the application of PCA in our study has allowed us to disentangle post-movement activity from the SPN itself. Indeed, one of the factor combinations we observed reflected this post-movement activity and was more negative contralateral to the response hand, while the factor combination that we argue reflects the SPN was maximal over the right hemisphere and did not reliably differ as a function of response hand.

**4.4. Reward Consumption**

Previous MID research has had issues observing reward consumption ERPs. Even when these ERPs are observed, their modulation is inconsistent with the wide literature on reward consumption (e.g., Broyd et al., 2012; Novak & Foti, 2015). In this study, the novel application of PCA to feedback-locked ERPs in the MID task successfully separated and quantified the RewP and fb-P3. This is despite the substantial temporal and spatial overlap that has contributed to equivocal effects reported in the past. Consistent with research applying PCA to a non-MID task (Foti & Hajcak, 2009), the RewP had a frontocentral distribution and was greater to feedback signalling gains than losses. RewP amplitudes on probable win and probable loss trials did not differ, suggesting incentive cues do not modulate RewP presentation.

The RewPs independence from incentive and expectation effects is important given the results of previous MID research. As pointed out by Novak and Foti (2015), loss and break-even

feedback on MID tasks are typically less frequent than reward. This element of their design confounds the effects of feedback type with stimulus probability making quantification of the RewP difficult. Importantly, in the present study, probable gain and probable loss trials had opposite feedback valence probabilities and produced, as expected, identical RewP amplitudes.

This finding is inconsistent with some studies that have examined feedback processing from a reinforcement learning perspective (Holroyd, Krigolson, & Lee, 2011; Ichikawa, Siegle, Dombrovski, & Ohira, 2010; Pfabigan, Alexopoulos, Bauer, & Sailer, 2011). In these studies, the RewP (or FRN) has typically been investigated as an electrocortical manifestation of reward prediction errors, with more negative amplitudes being observed for loss feedback than for gain feedback, and the difference between feedback valences being enhanced by greater reward expectancies (Bellebaum & Daum, 2008; Bellebaum et al., 2010), and when outcomes differ from expectations (Pfabigan et al., 2015). However, these effects have not always been consistently observed, with several studies finding that the FRN is sensitive to binary differences in outcome (e.g., reward vs nonreward) and not expectancy and expectancy violation (Highsmith, Wuensch, Tran, Stephenson, & Everhart, 2016). The findings of the present study are in broad agreement with research that that interprets the RewP as a binary index of desired versus undesired outcomes, rather than a measure of expectancy violation and prediction error per se. If a prediction error or expectancy violation effect was observed, then RewP amplitudes would be larger on trials where participants had viewed incentive cues that indicated a likely counter-factual outcome. It is important to note that because we did not vary reward probabilities throughout our task on a block by block or trial by trial basis it is difficult to draw any definitive conclusions with respect to which model of the RewP is best supported.

In contrast to the RewP, the fb-P3 tracked expectancy violation. Greater amplitudes were observed when feedback signalled losses on probable gain trials and gains on probable loss trials. This pattern of results is consistent with the effects reported in previous studies where feedback events that are seen as infrequent or important are associated with greater fb-P3 amplitudes. Importantly, our results suggest that application of PCA to reward consumption ERPs in the MID task can measure both the valence aspects of the RewP, and expectance and salience aspects of the fb-P3.

We also observed for the first time a very late negativity distributed over the left frontal pole. There is currently no existing explanation for this effect in the literature on reward consumption. While this component could potentially reflect an SPN to the offset of feedback stimuli, there is limited evidence that this is the case. Firstly, while the time-course and polarity of the factor combination is consistent with the SPN, it's topology is not. Secondly, there is limited

evidence of the SPN being produced prior to feedback offset in other studies using the MID task, or in studies using gambling tasks (e.g., Foti & Hajack, 2012). As the negativity is sensitive to the participant's response hand, especially in the context of reward feedback, we speculate this component may be associated with encoding and updating some form of successful response prototype. The greater negativity observed for right-hand responses might be a function of hand dominance as all the participants in our sample were right-handed. Similar fronto-polar negativities have been observed in prospective memory tasks when individuals successfully encode the intention to act in the near future, although these effects have only been observed following viewing instruction stimuli, rather than following feedback stimuli (West, 2011; West & Moore, 2002). An alternative explanation could be that this factor combination also reflects processes similar to an eligibility trace, albeit at a different stage of processing (Doya, 2008). Further research is required to reproduce the presentation of this component and establish what feature or features of reward consumption it is associated with.

Lastly, while some studies (e.g., Pornpattananangkul & Nusslock, 2015) have reported LPP modulation to feedback stimuli in the MID task, we did not observe this in the present study.

### 4.5. Conclusion

The MID task offers a promising experimental approach to assessing and modelling reward processing and consumption. Results from our study show that the dynamics of reward anticipation can be parsed into discriminable stages that are differentially modulated by probable gain, probable loss, and break-even cues. In addition, use of PCA on feedback-locked ERPs can allow researchers to assess both the RewP and fb-P3 independently in the MID task. We also observe for the first time a late negativity distributed over the left frontal pole in response to task feedback. This new reward consumption component appears to be sensitive to response hand, especially in the context of monetary gain. Further investigation is required to understand the exact functions that these components signify in the context of reward anticipation and consumption.

## References


Angus, D. J., Kemkes, K., Schutter, D. J. L. G., & Harmon-Jones, E. (2015). Anger is associated with reward-related electrocortical activity: Evidence from the reward positivity. *Psychophysiology, 52*(10), 1271-1280. doi:10.1111/psyp.12460

Balconi, M., & Crivelli, D. (2010). FRN and P300 ERP effect modulation in response to feedback sensitivity: The contribution of punishment-reward system (BIS/BAS) and Behaviour Identification of action. *Neuroscience Research, 66*(2), 162-172. doi:10.1016/j.neures.2009.10.011

Baskin-Sommers, A. R., & Foti, D. (2015). Abnormal reward functioning across substance use disorders and major depressive disorder: considering reward as a transdiagnostic mechanism. *International Journal of Psychophysiology, 98*(2), 227-239. doi: 10.1016/j.ijpsycho.2015.01.011

Becker, M. P. I., Nitsch, A. M., Miltner, W. H. R., & Straube, T. (2014). A single-trial estimation of the feedback-related negativity and its relation to BOLD responses in a time-estimation task. *The Journal of Neuroscience, 34*(8), 3005-3012. doi: 10.1523/JNEUROSCI.3684-13.2014

Bellebaum, C., & Daum, I. (2008). Learning-related changes in reward expectancy are reflected in the feedback-related negativity. *European Journal of Neuroscience, 27*(7), 1823-1835. doi:10.1111/j.1460-9568.2008.06138.x

Bellebaum, C., Kobza, S., Thiele, S., & Daum, I. (2010). It was not my fault: Event-related brain potentials in active and observational learning from feedback. *Cerebral Cortex, 20*(12), 2874-2883. doi:10.1093/cercor/bhq038

Bellebaum, C., Polezzi, D., & Daum, I. (2010). It is less than you expected: the feedback-related negativity reflects violations of reward magnitude expectations. *Neuropsychologia, 48*(11), 3343-3350. doi:10.1016/j.neuropsychologia.2010.07.023

Berridge, K. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology, 191*(3), 391-431. doi:10.1007/s00213-006-0578-x

Berridge, K., & Kringelbach, Morten L. (2015). Pleasure Systems in the Brain. *Neuron, 86*(3), 646-664. doi:10.1016/j.neuron.2015.02.018

Berridge, K., Robinson, T. E., & Aldridge, J. W. (2009). Dissecting components of reward: 'liking', 'wanting', and learning. *Current opinion in pharmacology, 9*(1), 65-73. doi:10.1016/j.coph.2008.12.014

Broyd, S. J., Richards, H. J., Helps, S. K., Chronaki, G., Bamford, S., & Sonuga-Barke, E. J. S. (2012). An electrophysiological monetary incentive delay (e-MID) task: A way to decompose the different components of neural response to positive and negative monetary reinforcement. *Journal of Neuroscience Methods, 209*(1), 40-49. doi:10.1016/j.jneumeth.2012.05.015

Brunia, C. H. M. (1988). Movement and stimulus preceding negativity. *Biological Psychology, 26*(1–3), 165-178. doi:10.1016/0301-0511(88)90018-X

Brunia, C. H. M., Hackley, S. a., van Boxtel, G. J. M., Kotani, Y., & Ohgami, Y. (2011). Waiting to perceive: Reward or punishment? *Clinical Neurophysiology, 122*(5), 858-868. doi:10.1016/j.clinph.2010.12.039

Carlson, J. M., Foti, D., Harmon-Jones, E., & Proudfit, G. H. (2015). Midbrain volume predicts fMRI and ERP measures of reward reactivity. *Brain Struct Funct, 220*(3), 1861-1866. doi:10.1007/s00429-014-0725-9

Carlson, J. M., Foti, D., Mujica-Parodi, L. R., Harmon-Jones, E., & Hajcak, G. (2011). Ventral striatal and medial prefrontal BOLD activation is correlated with reward-related electrocortical activity: A combined ERP and fMRI study. *NeuroImage, 57*(4), 1608-1616. doi:10.1016/j.neuroimage.2011.05.037

Castro, D. C., & Berridge, K. (2014). Opioid hedonic hotspot in nucleus accumbens shell: mu, delta, and kappa maps for enhancement of sweetness "liking" and "wanting". The *Journal of Neuroscience, 34*(12), 4239-4250. doi:10.1523/JNEUROSCI.4458-13.2014

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*(2), 245-276. doi:10.1207/s15327906mbr0102_10

Damen, E. J. P., & Brunia, C. H. M. (1994). Is a stimulus conveying task-relevant information a sufficient condition to elicit a stimulus-preceding negativity? *Psychophysiology, 31*(2), 129-139. doi:10.1111/j.1469-8986.1994.tb01033.x

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9-21. doi:10.1016/j.jneumeth.2003.10.009

Delplanque, S., Silvert, L., Hot, P., Rigoulot, S., & Sequeira, H. (2006). Arousal and valence effects on event-related P3a and P3b during emotional categorization. *International Journal of Psychophysiology, 60*(3), 315-322. doi:10.1016/j.ijpsycho.2005.06.006

Dien, J. (2010a). The ERP PCA Toolkit: An open source program for advanced statistical analysis of event-related potential data. *Journal of Neuroscience Methods, 187*(1), 138-145. doi:10.1016/j.jneumeth.2009.12.009

Dien, J. (2010b). Evaluating two-step PCA of ERP data with Geomin, Infomax, Oblimin, Promax, and Varimax rotations. *Psychophysiology, 47*(1), 170-183. doi:10.1111/j.1469-8986.2009.00885.x

Dien, J., Franklin, M. S., & May, C. J. (2006). Is "Blank" a suitable neutral prime for event-related potential experiments? *Brain Lang, 97*(1), 91-101. doi:10.1016/j.bandl.2005.08.002

Dien, J., & Frishkoff, G. A. (2005). Principal components analysis of event-related potential datasets. In T. Handy (Ed.), *Event-related potentials: a methods handbook*. Cambridge, MA: MIT Press.

Dien, J., Khoe, W., & Mangun, G. R. (2007). Evaluation of PCA and ICA of simulated ERPs: Promax vs. infomax rotations. *Human Brain Mapping, 28*(8), 742-763. doi:10.1002/hbm.20304

Doñamayor, N., Schoenfeld, M. A., & Münte, T. F. (2012). Magneto- and electroencephalographic manifestations of reward anticipation and delivery. *NeuroImage, 62*(1), 17-29. doi:10.1016/j.neuroimage.2012.04.038

Doya, K. (2008). Modulators of decision making. *Nature Neuroscience, 11*(4), 410-416. doi:10.1038/nn2077

Foti, D., & Hajcak, G. (2009). Depression and reduced sensitivity to non-rewards versus rewards: Evidence from event-related potentials. *Biological Psychology, 81*(1), 1-8. doi:10.1016/j.biopsycho.2008.12.004

Foti, D., & Hajcak, G. (2010). State sadness reduces neural sensitivity to nonrewards versus rewards. *NeuroReport, 21*(2), 143-147. doi:10.1097/WNR.0b013e3283356448

Foti, D., & Hajcak, G. (2012). Genetic variation in dopamine moderates neural response during reward anticipation and delivery: Evidence from event-related potentials. *Psychophysiology, 49*(5), 617-626. doi:10.1111/j.1469-8986.2011.01343.x

Foti, D., Hajcak, G., & Dien, J. (2009). Differentiating neural responses to emotional pictures: Evidence from temporal-spatial PCA. *Psychophysiology, 46*(3), 521-530. doi:10.1111/j.1469-8986.2009.00796.x

Foti, D., Weinberg, A., Bernat, E., & Proudfit, G. H. (2015). Anterior cingulate activity to monetary loss and basal ganglia activity to monetary gain uniquely contribute to the feedback negativity. *Clinical Neurophysiology, 126*(7), 1338-1347. doi:10.1016/j.clinph.2014.08.025

Gehring, W. J., & Willoughby, A. R. (2002). The Medial Frontal Cortex and the Rapid Processing of Monetary Gains and Losses. *Science, 295*(5563), 2279-2282. doi:10.1126/science.1066893

Goldstein, R. Z., Cottone, L. A., Jia, Z., Maloney, T., Volkow, N. D., & Squires, N. K. (2006). The effect of graded monetary reward on cognitive event-related potentials and behavior in young healthy adults. *International Journal of Psychophysiology, 62*(2), 272-279. doi:10.1016/j.ijpsycho.2006.05.006

Haber, S. N., & Knutson, B. (2009). The Reward Circuit: Linking Primate Anatomy and Human Imaging. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology, 35*(1), 1-23. doi:10.1038/npp.2009.129

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology, 71*(2), 148-154. doi:10.1016/j.biopsycho.2005.04.001

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2007). It's worse than you thought: The feedback negativity and violations of reward prediction in gambling tasks. *Psychophysiology, 44*(6), 905-912. doi:10.1111/j.1469-8986.2007.00567.x

Harmon-Jones, E., Harmon-Jones, C., & Price, T. F. (2013). What is approach motivation? *Emotion Review, 5,* 291-295. doi: 10.1177/1754073913477509

Hauser, T. U., Iannaccone, R., Stämpfli, P., Drechsler, R., Brandeis, D., Walitza, S., & Brem, S. (2014). The feedback-related negativity (FRN) revisited: New insights into the localization, meaning and network organization. *NeuroImage, 84*(0), 159-168. doi:10.1016/j.neuroimage.2013.08.028

Highsmith, J. M., Wuensch, K. L., Tran, T., Stephenson, A. J., & Everhart, D. E. (2016). It's not what you expect: feedback negativity is independent of reward expectation and affective responsivity in a non-probabilistic task. *Brain Informatics*, 1-13. doi:10.1007/s40708-016-0050-6

Hofmann, M. J., Kuchinke, L., Tamm, S., Võ, M. L. H., & Jacobs, A. M. (2009). Affective processing within 1/10th of a second: High arousal is necessary for early facilitative processing of negative but not positive words. *Cognitive, Affective, & Behavioral Neuroscience, 9*(4), 389-397. doi:10.3758/9.4.389

Holroyd, C. B., Krigolson, O. E., & Lee, S. (2011). Reward positivity elicited by predictive cues. *NeuroReport, 22*(5), 249-252. doi:10.1097/WNR.0b013e328345441d

Homberg, V., Grunewald, G., & Grunewald-Zuberbier, E. (1981). The variation of p300 amplitude in a money-winning paradigm in children. *Psychophysiology, 18*(3), 258-262. doi:10.1111/j.1469-8986.1981.tb03030.x

Ichikawa, N., Siegle, G. J., Dombrovski, A., & Ohira, H. (2010). Subjective and model-estimated reward prediction: Association with the feedback-related negativity (FRN) and reward

prediction error in a reinforcement learning task. *International Journal of Psychophysiology, 78*(3), 273-283. doi:10.1016/j.ijpsycho.2010.09.001

Johnson-Frey, S. H., Newman-Norlund, R., & Grafton, S. T. (2005). A distributed left hemisphere network active during planning of everyday tool use skills. *Cerebral Cortex, 15*(6), 681-695. doi:10.1093/cercor/bhh169

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology, 37*(02), 163-178. doi:10.1111/1469-8986.3720163

Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. *Psychophysiology, 40*(4), 586-596. doi:10.1111/1469-8986.00060

Keysers, C., Kaas, J. H., & Gazzola, V. (2010). Somatosensation in social perception. *Nature Reviews Neuroscience*, *11*(6), 417-428. doi:10.1038/nrn2833

Knutson, B., Fong, G. W., Adams, C. M., Varner, J. L., & Hommer, D. (2001). Dissociation of reward anticipation and outcome with event-related fMRI. *NeuroReport, 12*(17), 3683-3687. doi:10.1097/00001756-200112040-00016

Knutson, B., & Greer, S. M. (2008). Anticipatory affect: neural correlates and consequences for choice. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1511), 3771-3786. doi:10.1098/rstb.2008.0155

Knutson, B., Taylor, J., Kaufman, M., Peterson, R., & Glover, G. (2005). Distributed neural representation of expected value. *The Journal of Neuroscience, 25*(19), 4806-4812. doi:10.1523/JNEUROSCI.0642-05.2005

Knutson, B., Westdorp, A., Kaiser, E., & Hommer, D. (2000). FMRI Visualization of Brain Activity during a Monetary Incentive Delay Task. *NeuroImage, 12*(1), 20-27. doi:10.1006/nimg.2000.0593

Kotani, Y., Ohgami, Y., Yoshida, N., Kiryu, S., & Inoue, Y. (2017). Anticipation process of the human brain measured by stimulus-preceding negativity (SPN). *The Journal of Physical Fitness and Sports Medicine, 6*(1), 7-14. doi:10.7600/jpfsm.6.7

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An Open-Source Toolbox for the Analysis of Event-Related Potentials. *Frontiers in Human Neuroscience, 8*. doi:10.3389/fnhum.2014.00213

Luck, S. J. (2014). *An introduction to the event-related potential technique*: MIT Press, Cambridge.

Masaki, H., Yamazaki, K., & Hackley, S. (2010). Stimulus-preceding negativity is modulated by action-outcome contingency. *NeuroReport, 21*(4), 277-281. doi: 10.1097/WNR.0b013e3283360bc3

Mühlberger, C., Angus, D. J., Jonas, E., Harmon-Jones, C., & Harmon-Jones, E. (2017). Perceived control increases the reward positivity and stimulus preceding negativity. *Psychophysiology*, *54*, 310–322. doi:10.1111/psyp.12786

Müller, M., Rockstroh, B., Berg, P., Wagner, M., Elbert, T., & Makeig, S. (1994). SSR-modulation during slow cortical potentials. In C. Pantev, T. Elbert, & B. Lütkenhöner (Eds.), *Oscillatory Event-Related Brain Dynamics* (pp. 325-341). New York: Springer.

Novak, K. D., & Foti, D. (2015). Teasing apart the anticipatory and consummatory processing of monetary incentives: An event-related potential study of reward dynamics. *Psychophysiology, 52*, 1470-1482. doi:10.1111/psyp.12504

Ohgami, Y., Kotani, Y., Hiraku, S., Aihara, Y., & Ishii, M. (2004). Effects of reward and stimulus modality on stimulus-preceding negativity. *Psychophysiology, 41*(5), 729-738. doi:10.1111/j.1469-8986.2004.00203.x

Ohgami, Y., Kotani, Y., Tsukamoto, T., Omura, K., Inoue, Y., Aihara, Y., & Nakayama, M. (2006). Effects of monetary reward and punishment on stimulus-preceding negativity. *Psychophysiology, 43*(3), 227-236. doi:10.1111/j.1469-8986.2006.00396.x

Otten, L. J., Gaillard, a. W., & Wientjes, C. J. (1995). The relation between event-related brain potential, heart rate, and blood pressure responses in an S1-S2 paradigm. *Biological Psychology, 39*(2-3), 81-102. doi:10.1016/0301-0511(94)00969-5

Pfabigan, D. M., Alexopoulos, J., Bauer, H., & Sailer, U. (2011). Manipulation of feedback expectancy and valence induces negative and positive reward prediction error signals manifest in event-related brain potentials. *Psychophysiology, 48*(5), 656-664. doi: 10.1111/j.1469-8986.2010.01136.x

Pfabigan, D. M., Seidel, E. M., Sladky, R., Hahn, A., Paul, K., Grahl, A., . . . Lamm, C. (2014). P300 amplitude variation is related to ventral striatum BOLD response during gain and loss anticipation: An EEG and fMRI experiment. *NeuroImage, 96*, 12-21. doi:10.1016/j.neuroimage.2014.03.077

Pfabigan, D. M., Seidel, E.-M., Paul, K., Grahl, A., Sailer, U., Lanzenberger, R., . . . Lamm, C. (2015). Context-sensitivity of the feedback-related negativity for zero-value feedback outcomes. *Biological Psychology, 104*, 184-192. doi:10.1016/j.biopsycho.2014.12.007

Plichta, M. M., Wolf, I., Hohmann, S., Baumeister, S., Boecker, R., Schwarz, A. J., . . . Meyer, P. (2013). Simultaneous EEG and fMRI reveals a causally connected subcortical-cortical

network during reward anticipation. *The Journal of Neuroscience, 33*(36), 14526-14533. doi:10.1523/JNEUROSCI.0631-13.2013

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology, 118*(10), 2128-2148. doi:10.1016/j.clinph.2007.04.019

Pornpattananangkul, N., & Nusslock, R. (2015). Motivated to win: Relationship between anticipatory and outcome reward-related neural activity. *Brain and Cognition, 100*, 21-40. doi:10.1016/j.bandc.2015.09.002

Proudfit, G. H. (2015). The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology*, *52*(4), 449-459. doi:10.1111/psyp.12370

Proverbio, A. M., Burco, F., del Zotto, M., & Zani, A. (2004). Blue piglets? Electrophysiological evidence for the primacy of shape over color in object recognition. *Cognitive Brain Research, 18*(3), 288-300. doi:10.1016/j.cogbrainres.2003.10.020

Ramsey, S. E., & Finn, P. R. (1997). P300 from men with a family history of alcoholism under different incentive conditions. *Journal of Studies on Alcohol, 58*(6), 606-616. doi: 10.15288/jsa.1997.58.606

Santesso, D. L., Bogdan, R., Birk, J. L., Goetz, E. L., Holmes, A. J., & Pizzagalli, D. A. (2012). Neural responses to negative feedback are related to negative emotionality in healthy adults. *Social Cognitive and Affective Neuroscience, 7*(7), 794-803. doi:10.1093/scan/nsr054

Sobotka, S. S., Davidson, R. J., & Senulis, J. A. (1992). Anterior brain electrical asymmetries in response to reward and punishment. *Electroencephalography and Clinical Neurophysiology, 83*(4), 236-247. doi:

West, R. (2011). The temporal dynamics of prospective memory: A review of the ERP and prospective memory literature. *Neuropsychologia, 49*(8), 2233-2245. doi:10.1016/j.neuropsychologia.2010.12.028

West, R., & Moore, K. (2002). Adjustments of Cognitive Control in Younger and Older Adults. *Cortex, 41,* 570-581. doi:10.1016/S0010-9452(08)70197-7

Vignapiano, A., Mucci, A., Ford, J., Montefusco, V., Plescia, G. M., Bucci, P., & Galderisi, S. (2017). Reward anticipation and trait anhedonia: An electrophysiological investigation in subjects with schizophrenia. *Clinical Neurophysiology, 127*(4), 2149-2160. doi:http://dx.doi.org/10.1016/j.clinph.2016.01.006

Yeung, N., & Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *The Journal of Neuroscience, 24*(28), 6258-6264. doi:10.1523/JNEUROSCI.4537-03.2004

Yu, R., & Zhou, X. (2006). Brain potentials associated with outcome expectation and outcome evaluation. *NeuroReport, 17*(15), 1649-1653. doi:10.1097/01.wnr.0000236866.39328.1d

## Appendix 1

One factor combination appeared to reflect the CNV. This combination, TF1/SF4 had a fronto-central distribution and peaked between 1840 and 1844ms following the onset of cue stimuli (Fig A.1). While the topology and latency of this factor combination is consistent with previous reports of CNV modulation by incentive cues (Novak & Foti, 2015; Plichta et al., 2013), we did not observe reliable differences between probable gain, probable loss, and break-even cues. A robust ANOVA using a Bonferroni adjusted $p$ value threshold of 0.003125 indicated that there was no significant effect of incentive cue ($T_{WJt}/c(2.0,16.0) = 3.07$, $p = 0.082$), cued response hand ($T_{WJt}/c(1.0,18.0) = 12.53$, $p = 0.0044$), or an interaction between the two, $T_{WJt}/c(2.0,16.0) = 0.01$, $p = 0.98$.

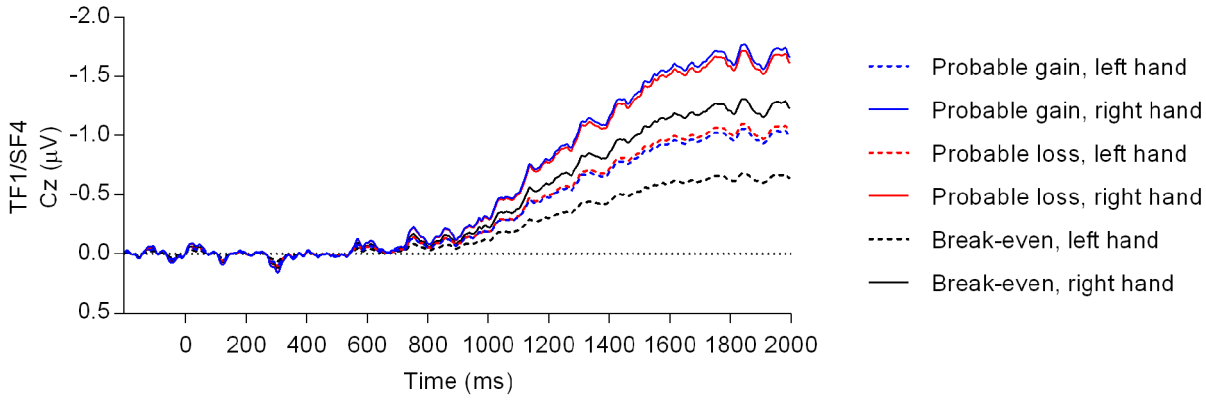


**Figure A1.** Cue locked waveforms for factor combination reflecting the CNV. Waveforms are presented in microvolt scale.

## Appendix 2

One factor combination appeared to reflect the SPN. This combination, TF1/SF1 had a fronto-central distribution and peaked between 4 ms before the onset of feedback stimuli (Fig A.2). Although the topology and latency of this factor combination is consistent with previous reports of SPN modulation by reward expectancy (Brunia, Hackley, van Boxtel, Kotani, & Ohgami, 2011; Ohgami, Kotani, Hiraku, Aihara, & Ishii, 2004; Ohgami et al., 2006; Pornpattananangkul & Nusslock, 2015), there were no reliable differences between probable gain, probable loss, and break-even cues. A robust ANOVA using a Bonferroni adjusted $p$ value threshold of 0.0055556 indicated that there was no significant effect of incentive cue ($T_{WJt}/c(2.0,16.0) = 1.85$, $p = 0.20$), cued response hand ($T_{WJt}/c(1.0,18.0) = 4.68$, $p = 0.048$), or an interaction between the two, $T_{WJt}/c(2.0,16.0) = 2.00$, $p = 0.20$.

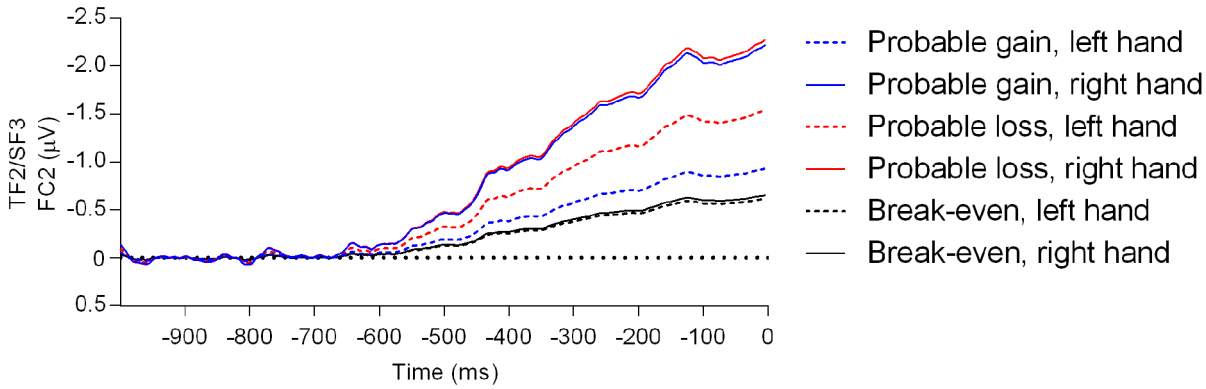


**Figure A2.** Feedback anticipation waveforms for factor combination reflecting the SPN. Waveforms are presented in microvolt scale.